Chapter 6

Incipient speciation in Lake Massoko, Tanzania

6.1 Introduction

6.1.1 Publication note

The work described in this chapter, except section 6.3, has been published as M. Malinsky et al., ‘Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake’. *Science*. 350, 1493-1498 (2015) [200].

6.1.2 Background

In the summer of 2011, my collaborators Martin Genner and George Turner conducted a survey of fish fauna in six crater lakes in the Rungwe District of Tanzania (Tables 6.1, B.1)[1]. In all six lakes, they found endemic haplochromine cichlids of the genus *Astatotilapia*, closely related to *Astatotilapia calliptera* (Figure 6.1A), a species widely-distributed in the rivers, streams and shallow lake margins of the region, including in Lake Malawi itself. Thus, the Rungwe District *Astatotilapia* are close relatives of the of Lake Malawi endemic haplochromine cichlids.

In Lake Massoko, the benthic zone in deep waters (~20-25m) is very dimly lit and populated by cichlids with phenotypes clearly different to those typical of shallow waters (~<5m) close to the shore (littoral). Deep-water males are dark blue-black, while most males collected from the shallow waters are yellow-green, similar to riverine

---

[1] Except for a brief mention of the presence of *Tilapia squamipinnins* and *Astatotilapia calliptera* in Lake Massoko by Ricardo [201], there were to our knowledge no published records about this fish fauna.

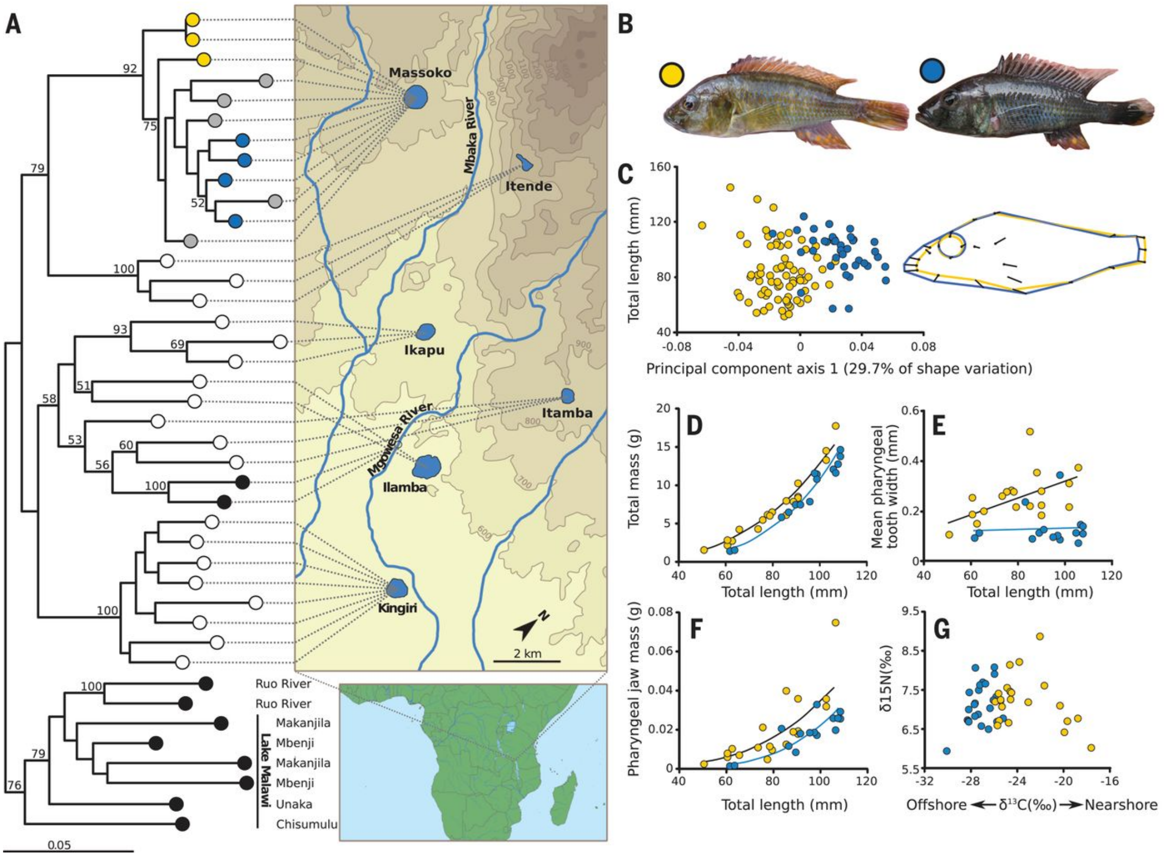


Fig. 6.1 **Cichlid radiation in the crater lakes of southern Tanzania. (A)** A phylogeny of the crater lake *Astatotilapia* based on reference-aligned RAD data (9,206SNPs). It demonstrates reciprocal monophyly between the populations in each lake except for Itamba, and close relationship to *A. calliptera* from rivers and from Lake Malawi. Within Lake Massoko, yellow symbols indicate the littoral morph, blue symbols indicate the benthic, and grey symbols denote small, phenotypically ambiguous, and thus unassigned individuals. Additional *Astatotilapia* individuals from other crater lakes are denoted by open circles. *A. calliptera* from rivers and Lake Malawi are denoted by black circles. Bootstrap values are displayed for nodes with >50% support. **(B)** Breeding males of the yellow littoral and blue benthic morphs of Lake Massoko. The symbols next to the photographs correspond to symbols used in (C-G). **(C-F)** Morphological divergence between the two morphs of Lake Massoko. Relative to the littoral, the benthic morph has relatively longer head and jaw **(C)**, lower body mass **(D)**, narrower 'papilliform' pharyngeal teeth **(E)**, and lighter lower pharyngeal jaws **(F)**. The benthic fish have stable isotope ratios that tend to be more depleted in C^{13} than the littoral, indicative of a more offshore-planktonic diet **(G)**.

forms (Figure 6.1B; Table 6.2). We also collected small (<65mm standard length) males that were not readily field-assigned to either ecomorph (Methods - section 6.8.1). The benthic and littoral morphs are reminiscent of the species pair of *Pundamilia* cichlids from Lake Victoria [89], but within a potentially simpler historical and geographical context. Lake Massoko is steep-sided, has a strong thermocline at ~15m, and an anoxic boundary at ~25m [202]. The estimated time of lake formation is ~50,000 years ago [203].

Table 6.1 **Location and geographical characteristics of crater lakes with haplochromine cichlid fauna in Rungwe District, Tanzania.** Data from [202], except Ikapu, estimated from Google Earth and own survey of depth.

Lake	Latitude	Longitude	Altitude (m)	Surface area (km^2)	Maximum depth (m)	Volume ($\times 10^6$m^3)
Kingiri	9°25’ S	33°51’ E	515	0.27	34	5.37
Ilamba	9°24’ S	33°50’ E	548	0.42	23	7.01
Ikapu	9°22’ S	33°48’ E	653	0.28	3	0.85
Itamba	9°21’ S	33°51’ E	821	0.12	18	0.69
Itende	9°19’ S	33°47’ E	1020	0.14	2	0.28
Massoko	9°20’ S	33°45’ E	845	0.38	37	8.91

Table 6.2 **Depth distribution of ecomorphs in Lake Massoko.** Based on collections using a variety of methods (Methods - section 6.8.1) in July-August and December 2014, and August 2015. There is a significant association between bottom depth and morph frequencies ($\chi^2_{4df} = 207.1$, P<0.001).

	0-5m	**5-10m**	**10-15m**	**15-20m**	**20-25m**	**Total**
Benthic	0	6	11	25	75	117
Littoral	98	54	15	21	0	188
Total	98	60	26	46	75	305
% Benthic	0	10	42.3	54.3	100	
% Littoral	100	90	57.7	45.7	0	

Ecomorph separation

To examine relationships between crater lake and riverine *A. calliptera* of southern Tanzania, Genner and Turner obtained restriction site associated DNA (RAD) [204] data from 30 fish from the Rungwe District, and 11 outgroup *Astatotilapia* from the broader Lake Malawi catchment (Figure 6.2, Table B.2). A maximum likelihood phylogeny constructed on the basis of these data demonstrates monophyly of all specimens from Lake Massoko (Figure 6.1A; Methods - section 6.8.2). Thus, the RAD phylogeny provides evidence that Massoko morphs might have evolved in primary sympatry, as previously proposed for crater lake cichlid radiations of Cameroon [205] and Nicaragua [206].

Morphological analyses of these two colour morphs revealed significant differences in head and body shape, body mass, the shape of pharyngeal teeth, and pharyngeal jaw mass (Figure 6.1C-F; Table 6.3; ANCOVA tests, all P<0.001). Genner and Turner also found significant differences in stable isotope ratios (Figure 6.1G; Table S5; ANCOVA test, P<0.001), indicative of dietary differences. Together these results demonstrate ecomorph separation and adaptation to different ecological environments

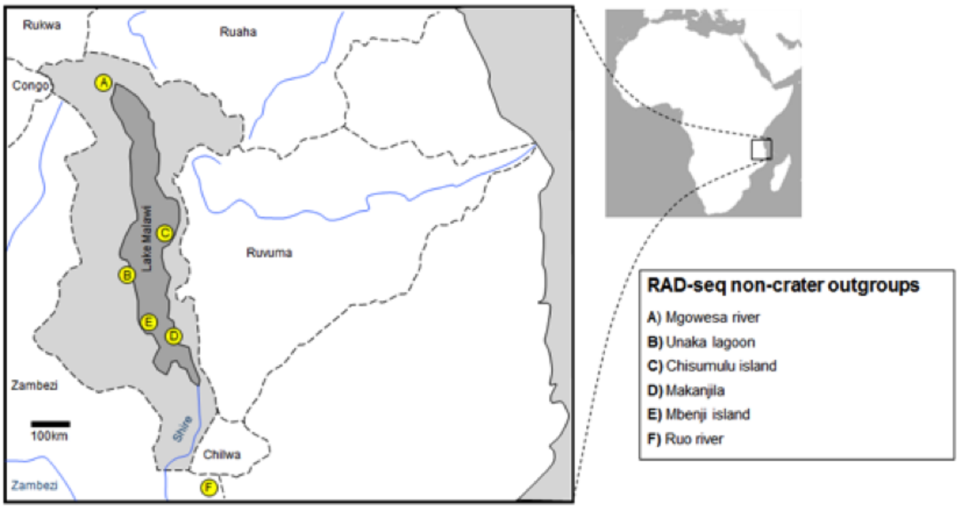


Fig. 6.2 **Collection sites of RAD sequenced non-crater-lake *Astatotilapia calliptera* for phylogenetic analysis.** Figure by M. Genner.

Table 6.3 **Results of morphological and stable isotope analysis.** Analysis of Covariance (ANCOVA) tests of morphological and stable isotope differences among benthic and littoral morphs. In each case total length (TL) was employed as a covariate.

	N benthic	N littoral	F TL	P TL	F ecomorph	P ecomorph
External morphology (PC1)	41	73	5.749	0.018	166.884	< 0.001
Body mass*	15	19	677.780	< 0.001	34.170	< 0.001
Pharyngeal jaw mass*	15	19	110.432	< 0.001	18.337	< 0.001
Pharyngeal jaw tooth width	15	19	4.037	0.053	25.121	< 0.001
Stable isotopes (δ^{13}C)	24	22	3.296	0.076	46.834	< 0.001
Stable isotopes (δ^{15}N)	24	22	0.516	0.476	0.636	0.430

*log10 transformed

The genomic causes and effects of divergent ecological selection during speciation are still poorly understood, in part because of the scarcity of well-characterised examples.

As I discussed in chapter 5, investigation of early stages of speciation in the large cichlid radiations of lakes Malawi, Tanganyika, and Victoria has been hampered by the complexity of those radiations, leading to difficulties in identifying sister species relationships, in reconstructing past geographical situations, and in controlling for possible introgression from non-sister taxa.

Therefore, while also working on improving our knowledge of species relationships in the Lake Malawi radiation (chapter 5), I focussed during my PhD on using whole genome DNA sequence data for a detailed study of ecomorph divergence in the much simpler Lake Massoko system. I investigated the geographical basis of the divergence, tested the 'genomic islands' model of ecological speciation, and explored functional correlations between highly divergent genomic regions and key traits likely to be involved in speciation, including mate choice and visual pigment spectral sensitivities.

6.2 Whole-genome evidence of Massoko divergence

To study the genome-wide pattern of Massoko ecomorph divergence and to further clarify its geographical context, we obtained whole-genome sequence data at ~15X coverage for 6 individuals each of the yellow littoral and blue benthic ecomorphs and 16 additional *A. calliptera* from the wider Lake Malawi catchment (Fig. S1), supplemented by lower coverage (~6X) data from 87 specimens from Lake Massoko (25 littoral, 32 benthic, and 30 small unassigned) and 30 individuals from Lake Itamba. This whole genome data has been described in chapter 3 (Figure 3.2; Table 3.3). The divergence from the *Metriaclima zebra* reference assembly was 0.2-0.3%, and variants were called at 4,755,448 sites (1.2-1.6 million sites per individual).

A maximum likelihood phylogeny built from whole genome sequence data confirmed reciprocal monophyly of *Astatotilapia* within Lakes Massoko and Itamba, and revealed the sister group of Massoko fish to be an *A. calliptera* population from the nearby Mbaka river (Figure 6.3A). All specimens of the benthic ecomorph formed a monophyletic clade derived from the littoral ecomorph (Figure 6.3A). Principal component analysis (PCA) showed strong population structure (Tracy-Widom statistics: $P<1\times10^{-12}$), with benthic and littoral individuals separated by the first eigenvector and forming separate clusters (Figure 6.3B). In contrast, within Lake Itamba, PCA did not reveal significant population structure (Tracy-Widom statistics: $P=0.11$). Individuals from Massoko that were not field-assigned to either of the ecomorphs did not form a monophyletic clade in the phylogeny (Figure 6.3A) or a distinct cluster in PCA (Figure 6.3B).

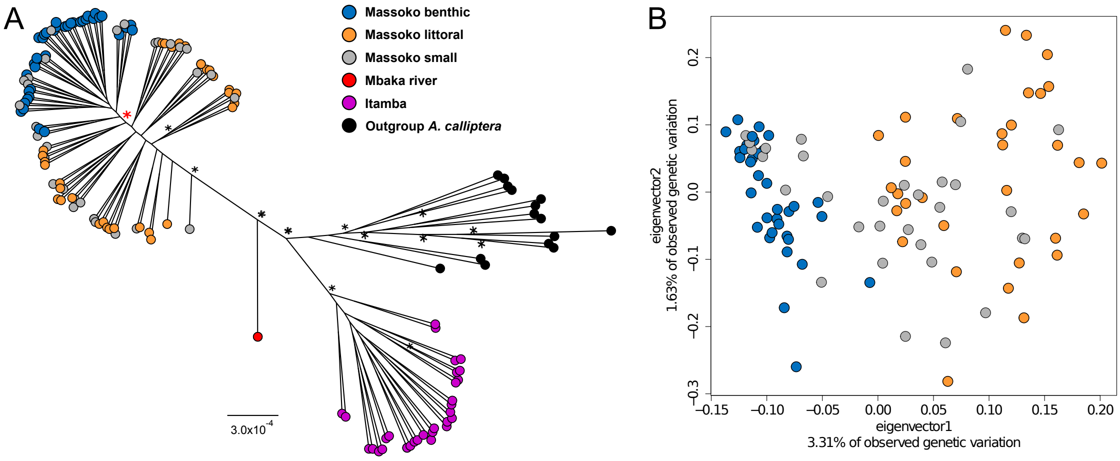


Fig. 6.3 **Lake Massoko divergence with whole genome sequence data. (A)** A maximum likelihood whole-genome phylogenetic tree. Black stars indicate nodes with 100% bootstrap support. The red star highlights the branch that separates all the Massoko benthic samples from the rest of the phylogeny (50% bootstrap support). **(B)** Principal Component Analysis of genetic variation within Lake Massoko.

I estimated individual ancestries for all Massoko specimens with the ADMIXTURE software [207] (Methods - section 6.8.3). Eleven of the 31 samples field-assigned as littoral, and 10 of the 30 unassigned individuals were identified as admixed (with the benthic gene pool), with admixture fraction >25% (Figure 6.4). On the other hand, no individuals identified as benthic were estimated to be admixed to the same extent. Therefore, recent gene flow may be biased from deep to shallow waters. We suggest that the remaining 20 unassigned samples represent sub-adult individuals of both benthic and littoral ecomorphs.

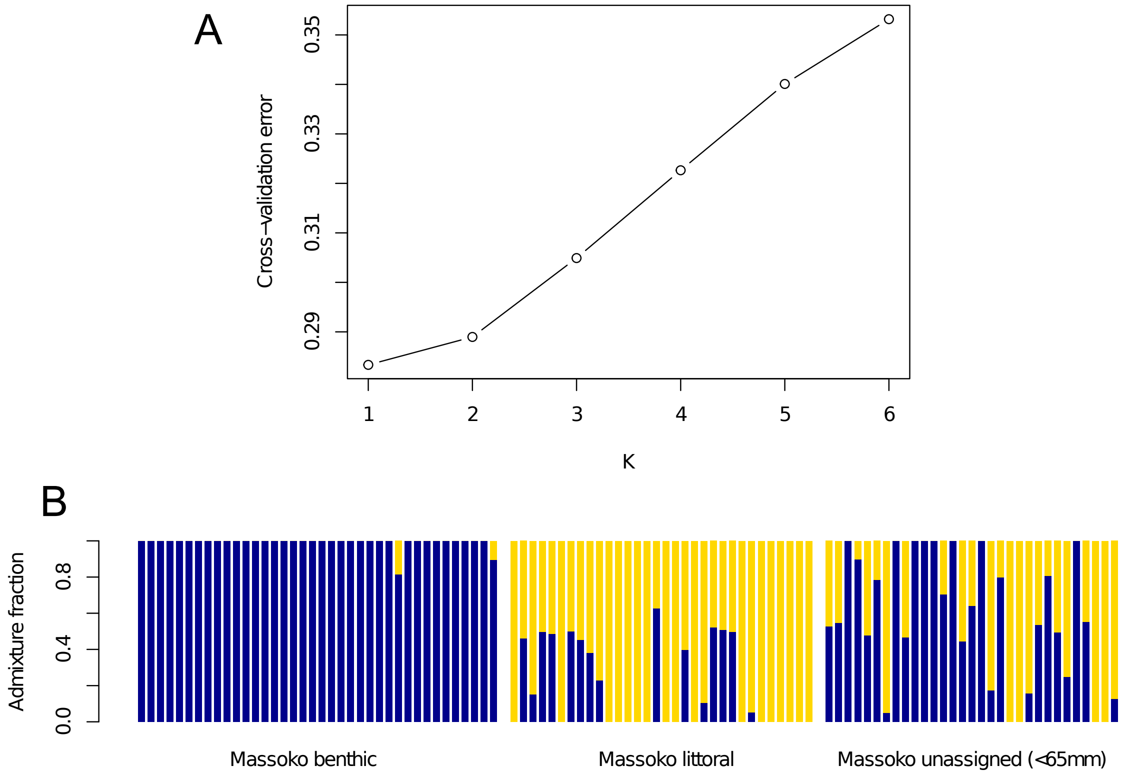


Fig. 6.4 **ADMIXTURE estimates of individual ancestries. (A)** ADMIXTURE cross-validation approach to choosing the `K` parameter - the postulated number of ancestral populations. The error estimates are based on 10-fold cross-validation. The lowest error is observed with `K=1`, suggesting that population differentiation between the ecomorphs is subtle [193] (Methods - section 6.8.3). **(B)** With two postulated ancestral populations (`K=2`), benthic individuals form a virtually homogenous group. Eleven of the samples field-assigned as littoral appear to be >25% admixed. The unassigned samples are a mixture of benthic, littoral, and admixed individuals.

Analysis of fine-scale genetic relationships with fineSTRUCTURE [193] supports the monophyly of the benthic ecomorph within the littoral, but also suggests that compared with the benthic population, the littoral population has greater co-ancestry with other *A. calliptera*; in particular with the Mbaka river sample (Figure 6.5).

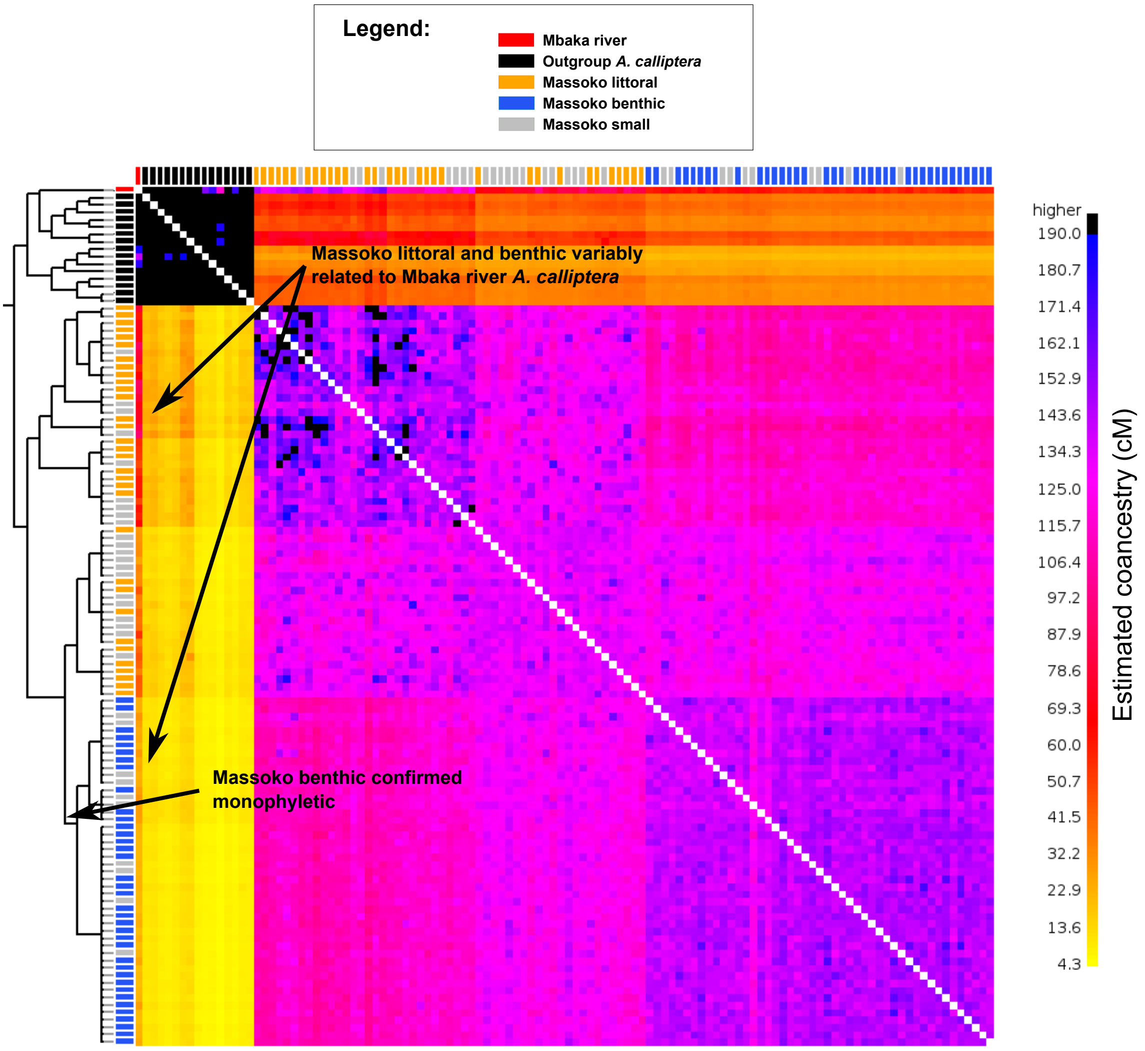


Fig. 6.5 **Massoko fineSTRUCTURE results.** Co-ancestry matrix with the tree showing inferred relationships between samples. Each tip and label correspond to an individual, with labels coloured according to the population/ecomorph as indicated in the legend. The results show tight clustering and monophyly of the benthic ecomorph, greater population structure within the littoral ecomorph, and a difference between the ecomorphs with respect to co-ancestry with Mbaka river *A. calliptera*, as indicated.

Therefore, I tested for evidence of secondary gene flow, as seen in cichlid populations from Cameroonian crater lakes [208]. Under the null hypothesis of no differential gene flow into Massoko, *A. calliptera* from Mbaka river should share derived alleles equally often with the littoral and with the benthic populations [190, 191]. Instead, we found an excess of shared derived alleles between *A. calliptera* from the Mbaka river and the littoral population, when compared with the benthic population (Patterson's D=1.1%; 4.86 SD from 0% or $P<5.8\times10^{-7}$) (Methods - section 6.8.3). The proportion of admixture f with Mbaka was estimated at 0.9±0.2%. However, this value is low, at a proportion that is approximately half of the Neanderthal introgression into non-African humans [190] and cross-coalescence rate analysis with MSMC [209] (Methods - section 6.8.3) indicates an average separation time of both Massoko ecomorphs from other *A. calliptera* samples (including Mbaka river) approximately ten times earlier than the split between the two ecomorphs (Figure 6.6). Thus, it is unlikely that a secondary invasion from the neighbouring river systems contributed to the divergence of the ecomorphs.

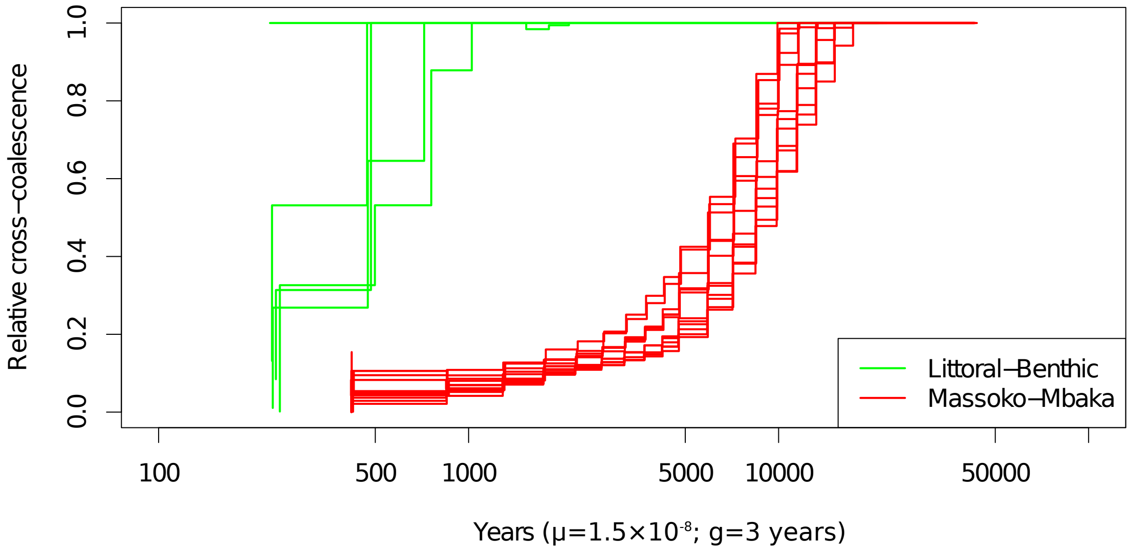


Fig. 6.6 **MSMC cross-coalescence between littoral and benthic ecomorphs (green) and between all Massoko high-coverage individuals and the sample from Mbaka river.** All Massoko individuals split from Mbaka approximately ten times earlier than any separation between benthic and littoral ecomorphs is observed. The time axis values assume: 1) average generation time g=three years and 2) per generation mutation rate $\mu=1.5\times10^{-8}$, making the assumption that the μ in cichlids is similar to μ estimated in human studies [15]. Direct estimate of μ in cichlids is not available.

6.3 Population size estimates

The number and frequency of heterozygous sites is a simple summary statistic estimating the level of genetic polymorhism in a population, and thus is indicative of long-term effective population size (N_e) over the past of order N_e generations [173]. Figure 6.7 shows heterozygosity in Itamba, Massoko, and in additional *A. calliptera* populations. Itamba individuals are considerably more heterozygous than fish from Massoko. This is interesting, because Itamba is smaller; it covers only 32% of the surface area, and has only 8% of the water volume of Massoko. Within Massoko, genome-wide average heterozygosity is lower in the benthic individuals compared with the littoral, consistent with benthic being the derived morph. There are also statistically significant differences between heterozygosity between the low-coverage and high-coverage samples, both in Massoko benthic (Welch two sample, two sided t-test: $p = 0.02$) and in Massoko littoral fish ($p = 4.6 \times 10^{-5}$). This result suggests that my variant calling pipeline under-calls heterozygous sites in low-coverage samples on average by approximately 3% in the benthic samples and 5% in the littoral samples. As already discussed in chapter 5, heterozygosity in the additional *A. calliptera* varies by more than an order of magnitude, reflecting their complex and disparate population histories.

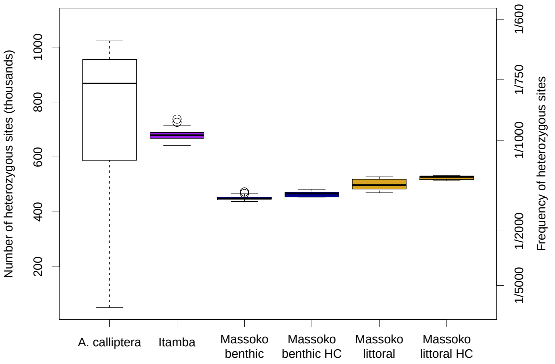


Fig. 6.7 **Heterozygosity in Itamba, Massoko, and additional *A. calliptera* populations** Lake Massoko individuals sequenced to high coverage (see Section 3.1) are shown separately, denoted by 'HC'.

The amount of linkage disequilibrium (LD) observed in a population can be used to estimate N_e over more recent history than sequence heterozygosity. LD between variants further apart from each other reflects more recent N_e than LD between variants that are closer together [210]. Figure 6.8 shows the decay of LD with distance in Itamba, and Massoko benthic and littoral ecomorphs. Interestingly, Massoko benthic fish have both the highest short distance LD, and the lowest long distance LD (beyond 42kb), suggesting a recent increase in N_e, compared to the there two populations. The Itamba samples, on the other hand, have lower short range LD (consistent with the high heterozygosity in Itamba), but have the highest level of LD for SNPs separated by more than 20kb. This pattern persists beyond the 100kb distance. The high level of long range LD in Itamba is evidence for low recent N_e.

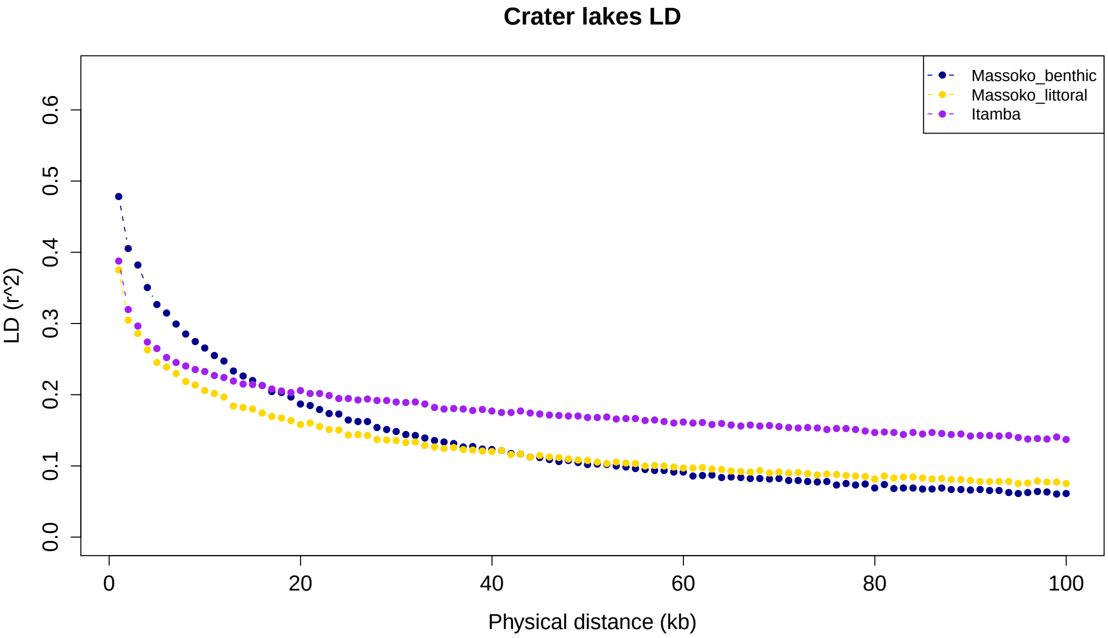


Fig. 6.8 **Decay of linkage disequilibrium in Itamba and Massoko ecomorphs** The figure shows average r^2 in 1kb windows.

MSMC estimates population size history

A more detailed picture of population size history may provide vital clues regarding the evolutionary history of the studied *Astatotilapia* populations. Therefore, I applied the Multiple sequentially Markovian coalescent (MSMC) model, developed by Schiffels and Durbin [209], to infer the history of N_e from the distribution of times since the most recent common ancestor between two alleles in an individual (2 haplotypes) or the first coalescence (common ancestor) for a pair of alleles among a set of four haplotypes. Using four haplotypes enables the algorithm to infer more recent N_e, because the first coalescence among the set of four haplotypes is typically more recent than when only two haplotypes are used.

Figure 6.9 shows MSMC estimates of N_e history for Massoko ecomorphs, Itamba, and *A. calliptera* from Mbaka river. For both Massoko ecomorphs, we see a pronounced drop in N_e (by approximately an order of magnitude), starting at 10,000 years ago, corresponding to the time of split from the Mbaka river population (Figure 6.6), and reaching a minimum at ~3,000. The drop could be related to a nearby volcanic eruption that threw out a layer of pumice that likely floated on the surface and affected the lake ecology, as suggested by M. Genner (pers. comm.) based on evidence from core samples. Consistent with the LD evidence, recent N_e increase in Massoko is more pronounced in the benthic ecomorph (starting at ~1,000 years ago, corresponding to the split time between the two ecomorphs: Figure 6.6). Also consistent with the LD evidence, the recent (~1,000 years ago) levels of N_e in Itamba are very low.

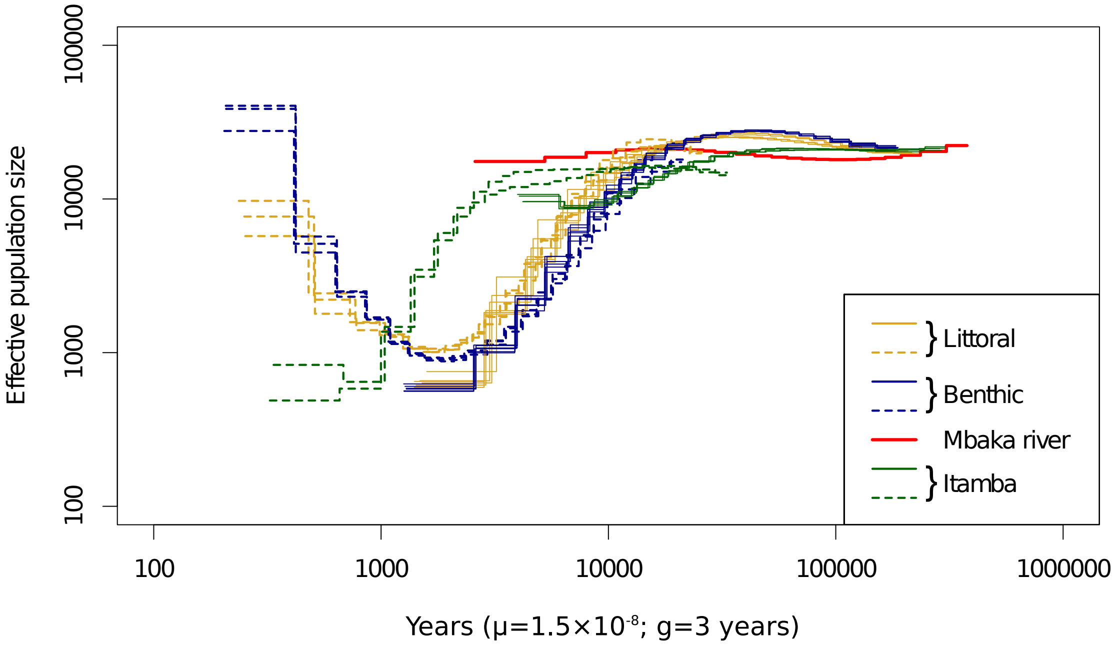


Fig. 6.9 **Inferred population size histories for Massoko ecomorphs, Itamba, and *A. calliptera* from Mbaka river.** Two haplotype inferences (solid lines) are based on six high coverage (HC) individuals each from the benthic and littoral ecomorphs, four Itamba individuals and the *A. calliptera* specimen collected from Mbaka river. Each solid line line corresponds to data from one individual. For the four-haplotype inference mode (inferring more recent N_e; dashed lines), I combined individuals into pairs (i.e. three pairs each for benthic and littoral ecomorphs; two pairs for Itamba). Thus each dashed line corresponds to data from two individuals. The four-haplotype inference mode could not be used for Mbaka river, as only one individual from Mbaka has been sequenced. Therefore, I do not have estimates of more recent population size history for Mbaka river. The time axis values assume: 1) average generation time g=three years and 2) per generation mutation rate $\mu = 1.5 \times 10^{-8}$, making the assumption that the μ in cichlids is similar to μ estimated in human studies [15]. Direct estimate of μ in cichlids is not available.

6.4 Islands of speciation

Interestingly, there are no fixed differences between Massoko benthic and littoral ecomorphs. Genome-wide divergence F_{ST} is 0.038, and almost half (47.6%) of the variable sites have zero F_{ST} (Table 6.4). Above the low background, a genome-wide F_{ST} profile shows clearly demarcated 'islands' of high differentiation (Figures 6.10A, 6.3C). For single sites, the maximum F_{ST} is 13.6 standard deviations (s.d.) above the mean, and 7,543 sites have F_{ST} over 6 s.d. above the mean. By contrast, comparisons of the combined Massoko population and Itamba population revealed a pattern of consistently high F_{ST} across the genome (Figure 6.10B). There are no statistical outliers (Figure 6.10C): not a single site has F_{ST} more than 3 s.d. away from the mean. Similar results were obtained when varying the window size to comprise 15, 50, 100, or 500 variants (Table 6.4; Figures B.1, B.2).

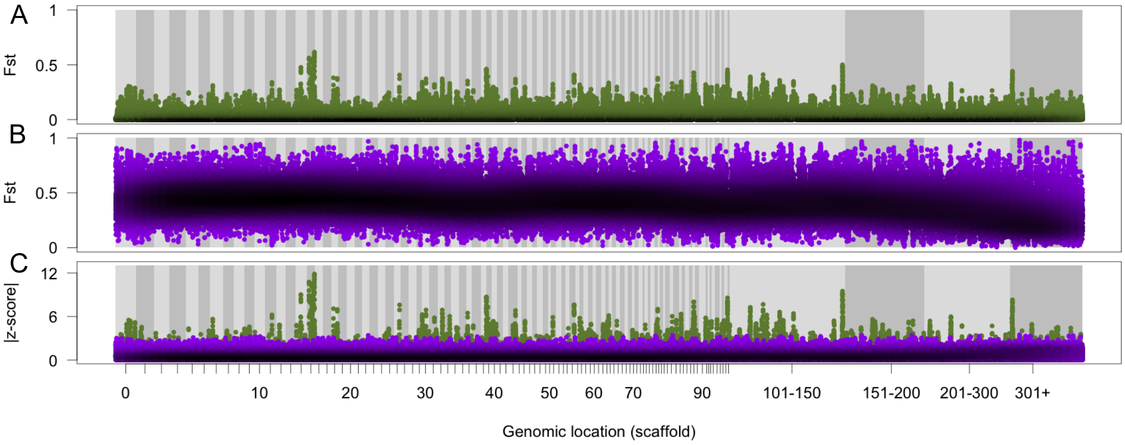


Fig. 6.10 **Genome-wide pattern of F_{ST} divergence in windows of 15 variants each.** Darker colour indicates greater density of datapoints. **(A)** Divergence between benthic and littoral ecomorphs within Massoko. **(B)** Divergence between combined Massoko and Itamba populations. **(C)** Absolute standard scores of Massoko-Itamba divergence (purple) overlaid on divergence between benthic and littoral ecomorphs (green).

Table 6.4 **A summary of sliding-window based F_{ST} calculations.**

Window size (variants)	**Average length (bp)**	**F_{ST} range**	**Median F_{ST}**	**Proportion with zero F_{ST}**	**95th percentile**	**99th percentile**
1	NA	0.00 - 0.72	0.003	0.476	0.126	0.247
15	5,369	0.00 - 0.66	0.016	0.258	0.134	0.24
50	17,839	0.00 - 0.62	0.018	0.208	0.129	0.231
100	35,455	0.00 - 0.60	0.019	0.171	0.126	0.225
500	174,390	0.00 - 0.46	0.024	0.064	0.115	0.197

A role of selection in generating genomic ‘islands’

To evaluate the extent to which the observed peaks of divergence could be explained by neutral processes, I used the coalescent simulator `ms` [211] to generate neutral (i.e. without selection) samples under two models of species formation: ‘Isolation after migration’ (IAM) and ‘Isolation with migration’ (IWM), as defined by Sousa and Hey [212] (Figure 6.11). Under both models, the divergence begins in the presence of gene-flow. Under the IWM model, gene-flow continues until the present, whereas under the IAM model it ceases at time T_1 (Figure 6.11).

Under both models I fixed the migration parameter `M=5` to simulate moderate bidirectional migration. The parameter `M` is is defined as $4N_0m$, where m is the fraction of each subpopulation made up of new migrants each generation. It was not intuitively clear to me how strong migration is implied by different values of `M`. Therefore, I calculated migration probabilities for a range values of values; these are presented in Table B.3. Next, under the IAM model I fixed the migration cessation time T_1 to equal half of the initial split time T_2. Finally, I fitted the split time parameter to obtain overall $F_{ST} = 0.038$ (over all the simulated sites), matching the overall F_{ST} divergence observed between the benthic and littoral ecomorphs (see Figure B.3).

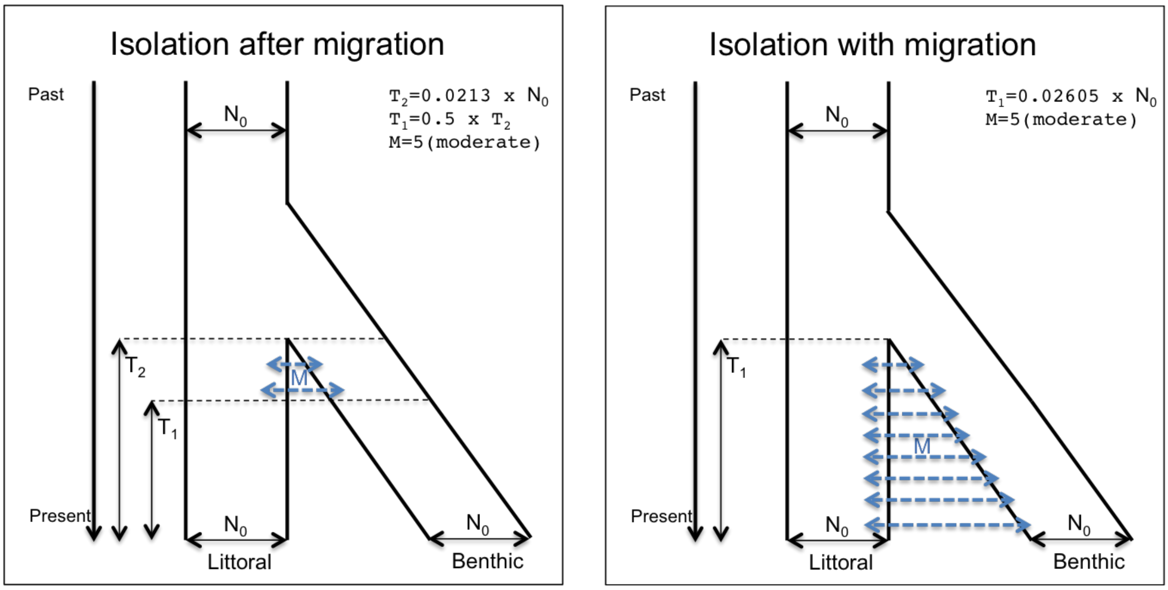


Fig. 6.11 **Models of species formation used for neutral coalescent simulations**

Comparing the distributions of F_{ST} values calculated from observed data (observed F_{ST}) and from the simulated data (simulated F_{ST}) using quantile-quantile plots (Figure 6.12) revealed that approximately the top 1% of observed F_{ST} values are higher than the corresponding simulated values, strengthening the evidence for the role of

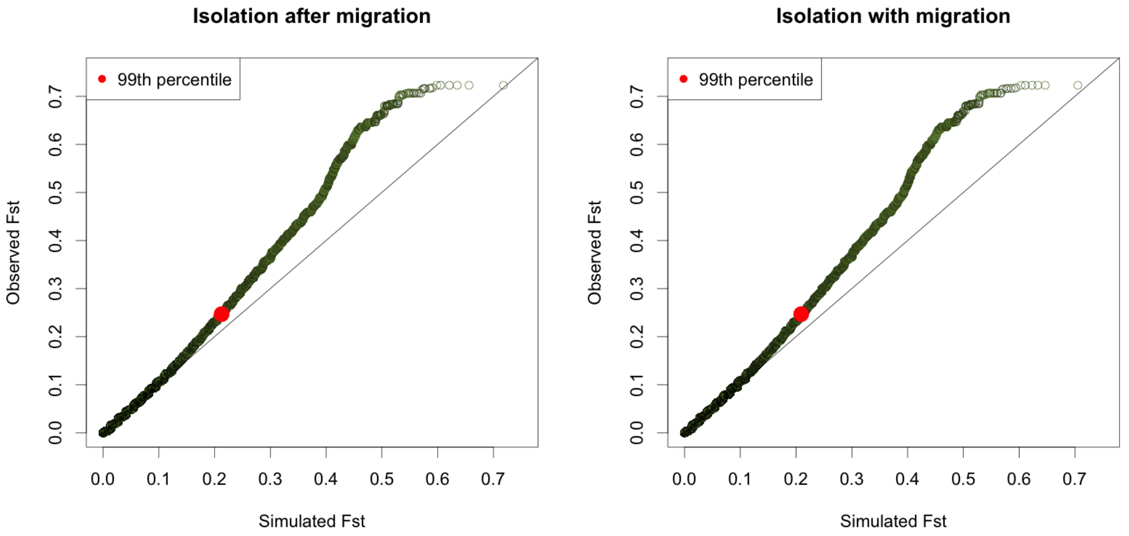


Fig. 6.12 **Comparing the distributions of observed and simulated $\mathbf{F}_{ST}$ values** Quantile-quantile plots comparing Massoko benthic-littoral F_{ST} observed in whole genome sequencing with F_{ST} from neutral coalescent simulations under two models of species formation. Darker colour indicates greater density of datapoints, and the position of the 99th percentile of both distributions is indicated in red. Very similar patterns are observed for simulations under both models.

divergent selection at these sites or at sites linked to them. The same conclusion can be drawn from simulations under both models.

To assess how the comparisons of observed and simulated F_{ST} values are affected by changes in simulated population sizes and the migration rates, I explored additional demographic scenarios for the IWM model. I simulated the IWM model with a period of reduced population size (bottleneck) around the split time between the two forms (Figure 6.13). Briefly, under this model, population sizes are reduced to $0.1{\times}N_0$ between the time points $T_1 = 0.5{\times}T_2$ and $T_3 = 1.5{\times}T_2$, where T_2 is the split time and N_0 is the current population size in both populations. For times further back in time (before $1.5{\times}T_2$), the population size is $0.8{\times}N_0$. These parameters were selected to approximately mimic population size changes inferred by MSMC (section 6.3). I also attempted to radically increase the migration parameter to `M=10`, `M=20` and `M=40`, significantly increasing the migration probabilities (Table B.3). As in all simulations, the split times (T_2) were fitted to match the overall observed F_{ST}.

The quantile-quantile plots (Figure 6.13) comparing F_{ST} from the additional simulations with F_{ST} values calculated from observed data show very little qualitative difference between the different demographic scenarios, confirming previous studies

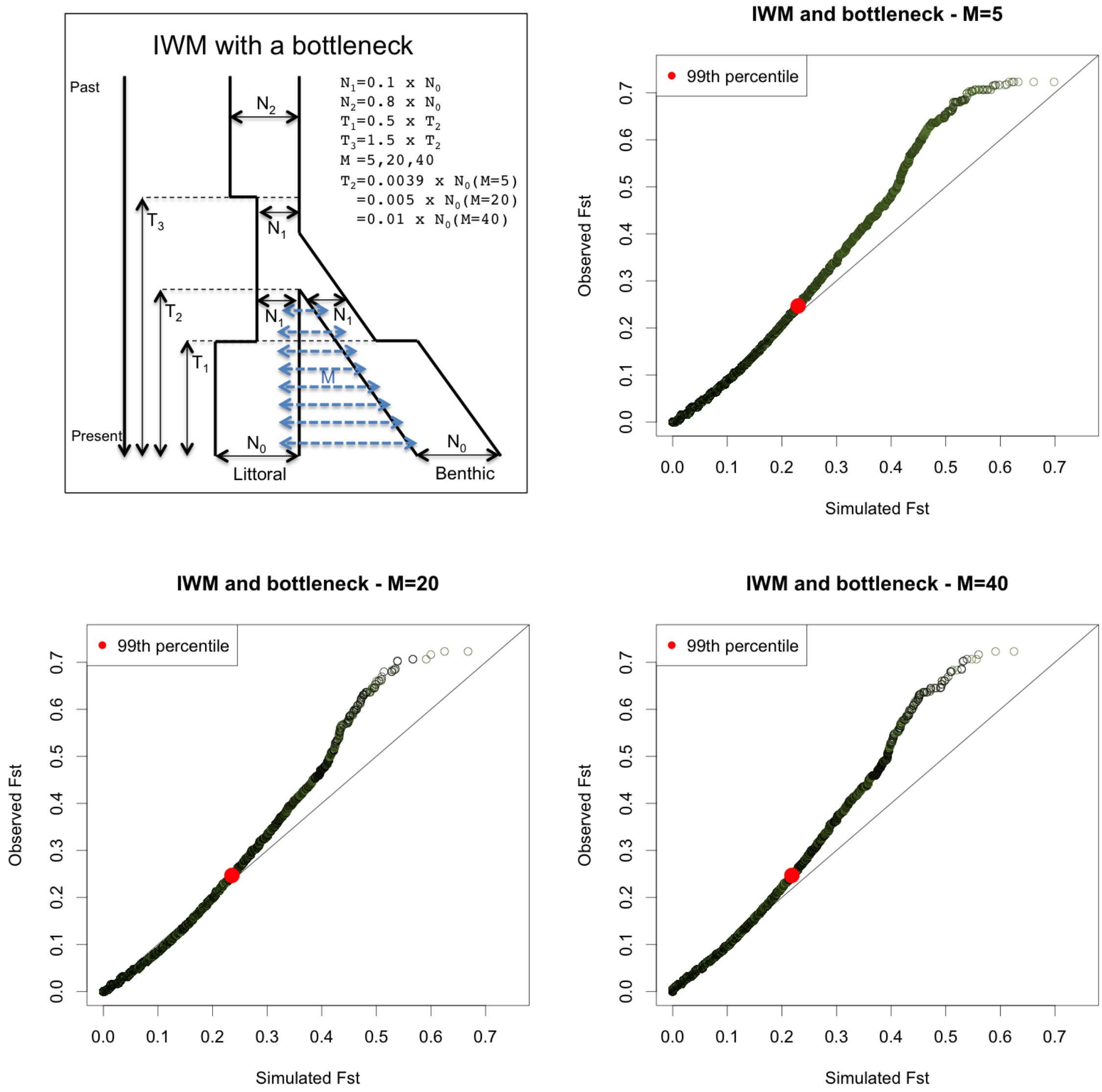


Fig. 6.13 **Comparing the distributions of observed and simulated F$_{ST}$ values for additional demographic models** The isolation with migration (IWM) model of species formation with a strong population bottleneck (top left) was used for neutral coalescent simulations, with a range of values for the migration parameter M. Quantile-quantile plots compare the distributions of simulated FST values under this model with observed benthic-littoral divergence (top-right and bottom). Darker colour indicates greater density of datapoints, and the position of the 99th percentile of both distributions is indicated in red. Similar patterns are observed for all simulations indicating approximately the top 1% of observed values are higher than corresponding simulated values.

and theoretical predictions asserting the distribution of F_{ST} values tends to be robust to demography [213]. Overall, these results suggests the following two conclusions: a) given an overall genome-wide level of F_{ST}, demography has a limited effect on the distribution of F_{ST} simulated under neutrality; b) the pattern of F_{ST} observed in the Massoko benthic-littoral divergence is very unlikely to have been generated by neutral processes alone. Specifically, the results are consistent with divergent selection acting on sites with approximately the top 1% of observed F_{ST} values (approximately $F_{ST} \geq 0.25$).

Loci underlying isolating traits in speciation with gene-flow

I identified genomic regions with observed benthic-littoral $F_{ST} \geq 0.25$ (i.e. with F_{ST} above maximum levels seen in neutral simulations) (Methods - section 6.8.3). For this I used windows of 15 variants each - providing a balance between fine genomic resolution and reducing stochastic variation by averaging over variants. Small gaps arising from brief dips of F_{ST} below the threshold were eliminated by merging regions within 10kb of one another. I found 344 such regions, with total length of 8.1Mb (~1% of the genome). Next, to focus on the more significant outliers, I narrowed the list down to a set of 98 highly diverged regions (HDRs) for further characterisation (Table B.4) by adding the requirement that at least one 10kb window must have reached $F_{ST} \geq 0.3$. The HDRs vary in length from 4.4kb to 285kb (median 36.1kb), with total length of 5.5Mb.

To investigate if the high F_{ST} divergence in HDRs is caused by reduced diversity or by reduced gene-flow, I calculated Nei's d_{XY}, an absolute measure of sequence divergence [66]. In contrast to studies re-examined by Cruickshank and Hahn [66], I found that, overall, d_{XY} in Massoko is significantly higher in the confirmed HDRs compared with the rest of the genome ($P<2.2\times10^{-16}$, two-tailed Mann-Whitney test; Figure 6.14A). Individually, 55 HDRs have d_{XY} above the 90th percentile of the genome-wide distribution. Post-split selective sweeps or other types of linked selection in the benthic or littoral populations would not be expected to generate such 'islands' of high differentiation in d_{XY} [66]. Therefore, these 55 regions (listed in Table 6.5) are the best candidates for loci underlying isolating barriers and being true 'islands of speciation'.

A key prediction of speciation with gene-flow models is that loci causing speciation should be located in relatively few linked clusters within the genome [71, 72]. As described by Cruikshank and Hahn: "With more than even a few islands that do not introgress because of selection, the number of recombinant individuals containing

Table 6.5 **Candidate 'islands of speciation'** The maximum (**`max`**) d_{XY}, π_{diff}, and F_{ST} values for each HDR, together with the quantile in the corresponding distributions (**F_x**). Linkage group (**`LG`**) assignment for each scaffold (**`sc`**) is based on a linkage map published in [214] - the **`?`** sign signifies that the scaffold could not be placed to a linkage group.

LG	**sc**	**start**	**end**	**max d_{XY}** $\times10^{-3}$	**F_x (max d_{XY})**	**max π_{diff}** $\times10^{-4}$	**F_x (max π_{diff})**	**max F_{ST}**	**F_x (max F_{ST})**
LG5	15	1934238	1967068	1.53	95.44%	5.33	97.62%	0.52	99.97%
	15	2912637	2961336	1.35	92.98%	1.83	75.74%	0.43	99.93%
	15	4641580	4880808	1.68	96.54%	2.92	88.81%	0.56	99.99%
	15	4907565	5049805	2.09	98.09%	5.23	97.47%	0.45	99.95%
	15	6705210	6818468	2.07	98.06%	11.2	99.84%	0.59	99.99%
	15	7208463	7304325	1.60	96.01%	5.12	97.31%	0.62	100.00%
	15	7507678	7592890	1.75	96.98%	4.86	96.82%	0.63	100.00%
	18	6702155	6728393	1.48	94.86%	9.39	99.66%	0.38	99.84%
LG7	0	11529594	11540402	1.92	97.66%	2.53	85.30%	0.31	99.65%
	0	11994849	12015103	1.55	95.62%	2.92	88.87%	0.31	99.61%
	32	4518067	4589712	1.80	97.21%	3.17	90.64%	0.39	99.88%
	32	4886114	4908718	1.66	96.41%	6.35	98.68%	0.32	99.67%
	99	355072	639642	1.41	94.03%	7.50	99.25%	0.49	99.97%
LG12	14	3582492	3609841	1.36	93.22%	6.43	98.74%	0.32	99.70%
	14	3661853	3697260	1.45	94.47%	6.01	98.41%	0.51	99.97%
	43	3655049	3696771	1.62	96.21%	1.41	67.26%	0.34	99.77%
LG20	57	46109	77869	1.36	93.19%	6.82	98.99%	0.33	99.72%
	108	814090	941030	2.04	97.96%	4.08	94.84%	0.39	99.87%
	164	0	113596	1.27	91.55%	5.26	97.51%	0.30	99.58%
	164	196412	276496	1.60	96.02%	7.91	99.40%	0.31	99.62%
LG23	30	183937	257768	1.62	96.19%	8.06	99.44%	0.35	99.79%
	30	797497	844062	1.31	92.39%	3.92	94.25%	0.32	99.71%
	31	4041909	4078026	1.47	94.77%	2.39	83.83%	0.32	99.68%
	82	2236206	2273645	1.75	96.99%	7.26	99.17%	0.36	99.80%
	88	819852	845401	1.27	91.42%	4.70	96.51%	0.34	99.76%
	88	1194601	1316288	1.50	95.08%	11.3	99.89%	0.41	99.90%
	88	1372483	1527476	1.86	97.48%	11.3	99.88%	0.46	99.95%
	88	1732907	1868455	1.38	93.55%	10.1	99.76%	0.35	99.79%
	95	1001404	1044619	1.35	93.01%	6.99	99.08%	0.33	99.71%
LG8	51	1450783	1493272	1.53	95.44%	3.93	94.31%	0.31	99.67%
	113	1062779	1122847	3.11	99.27%	5.08	97.25%	0.43	99.93%
	190	805453	832175	1.21	90.13%	1.08	58.40%	0.33	99.73%
LG3	126	420889	439080	3.78	99.53%	26.6	99.99%	0.32	99.69%
	186	693304	711017	1.38	93.60%	3.08	90.02%	0.30	99.60%
LG19	120	918534	962612	1.99	97.86%	4.35	95.64%	0.36	99.81%
	162	1227615	1263777	1.47	94.78%	1.49	69.10%	0.31	99.63%
?	39	465841	688825	2.37	98.59%	5.94	98.35%	0.47	99.96%
	39	2323506	2340670	1.22	90.40%	1.86	76.19%	0.31	99.66%
?	148	1458116	1644136	2.48	98.75%	10.3	99.74%	0.50	99.97%
	148	1669247	1754172	2.28	98.45%	6.05	98.45%	0.33	99.74%
LG18	6	2399603	2417150	1.46	94.65%	9.49	99.68%	0.30	99.61%
LG2	11	5426321	5452278	1.42	94.17%	2.97	89.21%	0.33	99.72%
LG13	26	5297874	5318387	1.59	95.97%	2.40	83.92%	0.30	99.59%
LG4	55	3423595	3500130	1.61	96.12%	4.76	96.63%	0.42	99.91%
LG11	64	55966	175700	1.42	94.08%	8.53	99.53%	0.34	99.75%
LG15	78	6039	59940	1.49	94.99%	2.28	82.59%	0.38	99.85%
LG14	84	2399084	2517997	1.40	93.81%	11.0	99.79%	0.38	99.85%
LG6	97	2188270	2212097	1.31	92.28%	2.25	82.16%	0.32	99.68%
LG9	229	470627	578767	1.87	97.53%	7.42	99.23%	0.39	99.86%
?	45	2785077	2828731	1.74	96.94%	3.24	91.09%	0.31	99.61%
?	91	129230	153938	1.62	96.20%	0.87	51.23%	0.34	99.75%
?	112	1966090	2014162	1.29	91.87%	7.81	99.38%	0.32	99.71%
?	114	1902474	2005892	2.10	98.13%	11.2	99.79%	0.32	99.70%
?	206	177202	290266	1.46	94.63%	8.66	99.56%	0.38	99.84%
?	304	0	70114	1.14	100.00%	18.9	99.98%	0.41	99.91%

the correct combination of parental alleles...becomes vanishingly small.” [66, p. 3151]. Instead of a large number of scattered islands, the theory predicts a smaller number of clusters that grow in size due to the ‘divergence hitchhiking’ process. We tested this prediction using a recently generated linkage map [214] and found that at least 27 out of the 55 putative speciation islands are co-localised on five linkage groups (LGs), with 26 of them clustered within their respective LGs (Figure 6.14B; Table 6.5). These potential speciation clusters extended for approximately 25cM on LG5, 40cM on LG7, 30cM on LG12, and 5cM on LG20 and 45cM on LG 23. In total, these regions account for under 7% of the genome, suggesting that divergence hitchhiking may play a role in shaping the observed pattern of genomic differentiation.

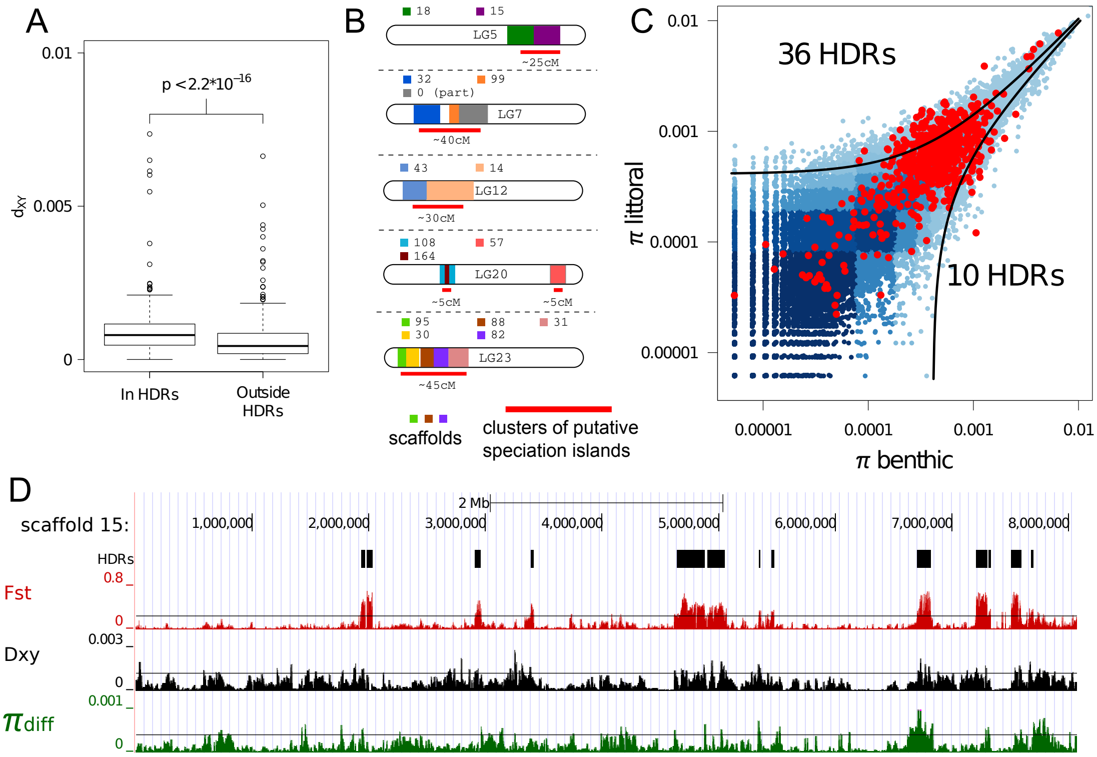


Fig. 6.14 **Islands of speciation between benthic and littoral ecomorphs. (A)** d_{XY} is significantly higher within HDRs ($p < 2.2\times10^{-16}$ one-tailed Mann-Whitney test), compared with the rest of the genome. **(B)** Clustering of putative speciation islands on five linkage groups. **(C)** Nucleotide diversity (π) within HDRs (red points) and outside HDRs (blue with shading corresponding to density). Each point corresponds to a 10kb window (therefore, there may be multiple points per HDR). 95% of observations lie between the two curves ($y=x\pm4.1\times10^{-4}$). Putative sweeps in the benthic ecomorph are in the top left corner and putative sweeps in the littoral in the bottom right corner. **(D)** Patterns of F_{ST}, d_{XY}, and π_{diff} in a speciation cluster on scaffold 15.

Although genomic islands within these clusters are often separated only by a few hundred kb, F_{ST} divergence between HDRs generally drops to background levels (see Figure 6.14D), with one exception on scaffold 88 where a broader 'continent' of divergence has formed (Figure 6.15).

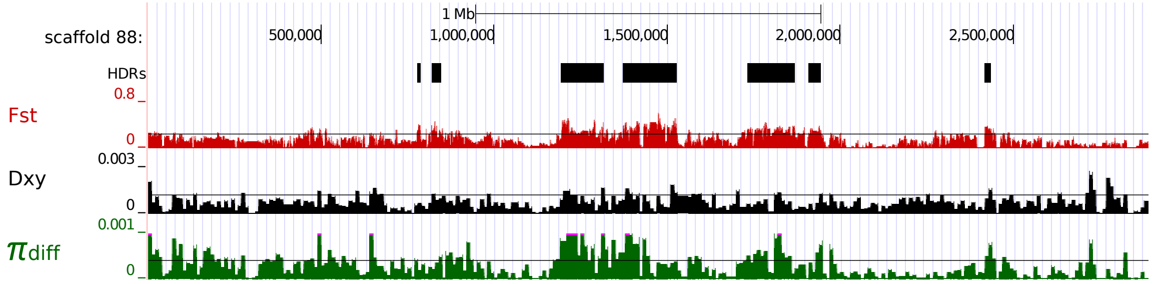


Fig. 6.15 **Patterns of F_{ST}, d_{XY}, and π_{diff} in a speciation cluster on scaffold 88.** In this region of the genome, F_{ST} appears to be elevated above or close to the 99th percentile (black line, representing maximum observed neutral divergence in simulations) over a large distance (2Mb), suggesting that divergence hitchhiking might have started forming a broader 'continent' of divergence.

Signals of adaptation

A reduced level of genetic polymorphism in one subpopulation may be indicative of a recent selective sweep. Overall, the magnitude of difference in nucleotide diversity (π) between benthic and littoral ecomorphs (π_{diff}) is significantly higher in the HDRs than in the rest of the genome ($P<2.2\times10^{-16}$, two-tailed Mann-Whitney test; Figure 6.16A) (Methods - section 6.8.3). Individually 46 HDRs have π_{diff} above the 95th percentile of the genome-wide distribution and are likely to have been under recent positive selection in one of the two ecomorphs. There is a significant overlap between HDRs with high d_{XY} (putative 'speciation islands') and HDRs with high π_{diff} (putative recent selective sweeps) - 35 of 55 high d_{XY} islands also have high π_{diff} (Figure 6.16B; $P=3\times10^{-5}$, hypergeometric test). On the other hand, the 11 putative sweeps that did not lead to elevated d_{XY} are indicative of adaptation not directly involved in reproductive isolation. Reduced nucleotide diversity in high π_{diff} regions, indicative of selective sweeps, was significantly more prevalent in the benthic ecomorph (36 of 46; $P<1.6\times10^{-4}$, two tailed Binomial test; Figures 6.14C, top left; 6.16C), consistent with the benthic ecomorph being derived and undergoing more extensive adaptation. Nevertheless, there are also a small number of strong outliers suggesting selective sweeps in the littoral ecomorph (Figure 6.14C, bottom right).

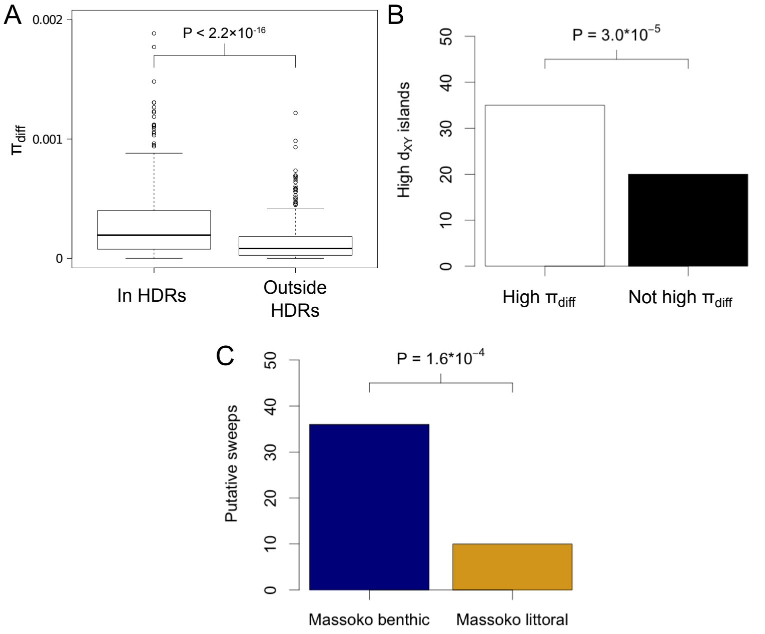


Fig. 6.16 **Characterisation of HDRs in terms of the magnitude of difference in nucleotide diversity between benthic and littoral ecomorphs (π_{diff}). (A)** π_{diff} is significantly higher within HDRs ($P < 2.2\times10^{-16}$, two-tailed Mann-Whitney test), compared with the rest of the genome. **(B)** The overlap between HDRs with high d_{XY} (putative 'speciation islands') and HDRs with high π_{diff} (putative recent selective sweeps) is significant. Thirty-five out of 55 high d_{XY} islands also have high π_{diff} ($P=3\times10^{-5}$, hypergeometric test). **(C)** Thirty-six out of the 46 putative selective sweeps are in the Massoko benthic morph, providing significant evidence that positive selection has been more prevalent in the benthic form ($P<1.6\times10^{-4}$, two tailed Binomial test).

Further support for sympatric divergence

We next tested whether the HDRs correlated with the signal of gene flow from the Mbaka river described above. Compared with the rest of the genome, the HDRs do not have elevated values of Patterson's D (P=0.22, two-tailed Mann-Whitney test; Figure 6.17C), nor elevated *f* statistics, which were recently proposed as an means by which one could identify introgressed loci [192] (Methods - section 6.8.3) (P=0.08, two-tailed Mann-Whitney test; Figure 6.17D). These results suggest that introgression from Mbaka river did not play a major role in generating the HDRs between the benthic and littoral ecomorphs within Lake Massoko (Figure 6.17), and strengthen the evidence that the ecomorph divergence has been happening within the lake.

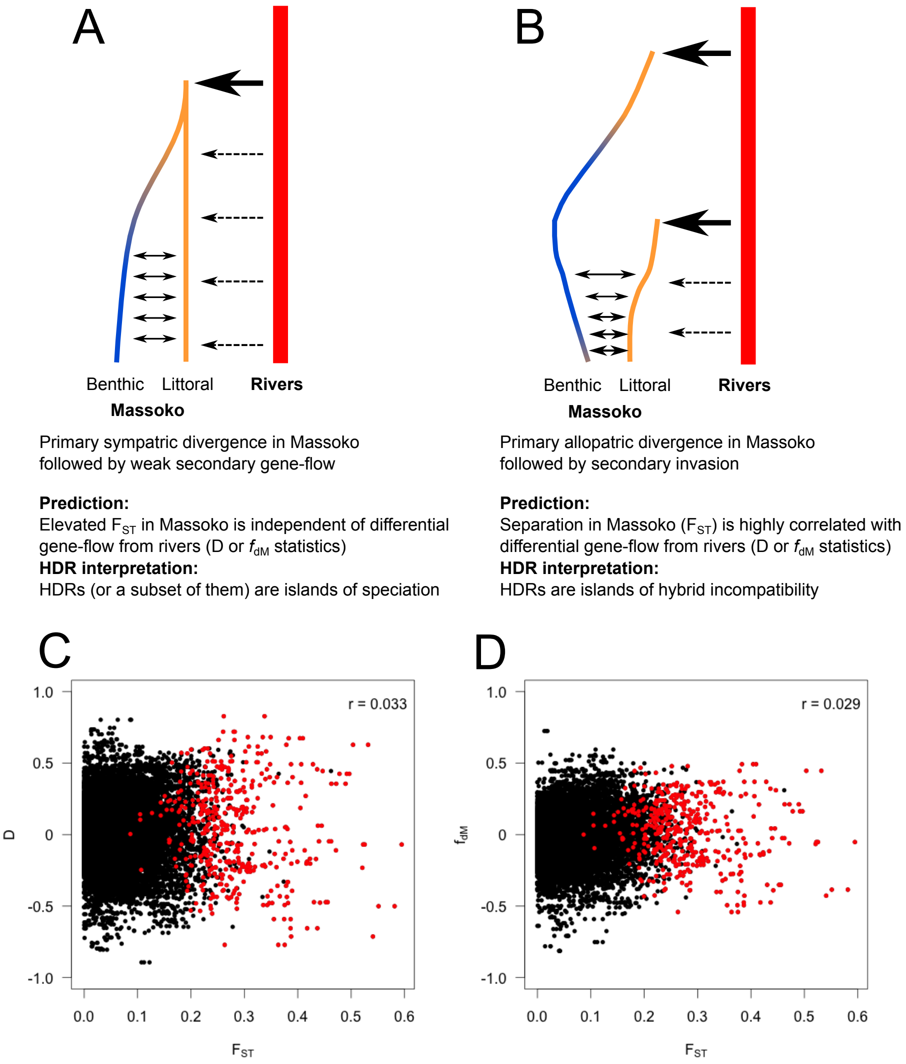


Fig. 6.17 **Evidence against allopatric (double-invasion) divergence between benthic and littoral ecomorphs.** **(A-B)** Two geographic models of divergence with predictions and interpretation of HDRs. **(A)** Primary sympatric divergence in Massoko. **(B)** Primary allopatric divergence. **(C-D)** F_{ST} between benthic and littoral ecomorphs is independent of levels of differential gene-flow measured by **(C)** Patterson's D, and **(D)** the f_{dM} statistic for locating introgressed regions. F_{ST} was averaged over windows with 15 variants, and D and f_{dM} were averaged over windows of 50 informative variants (i.e. variants with ABBA or BABA patterns). Values for windows within HDRs are shown in red, values for windows outside HDRs are shown in black. Pearson correlation coefficients r are displayed in the top right corner of each figure. The lack of correlation between either D or f_{dM} with F_{ST} is consistent with the predictions of the model of primary sympatric divergence.).

6.5 Divergent SNPs associated with mate choice

Many recently diverged taxa, particularly those not geographically isolated, show stronger pre-mating isolation than post-mating isolation [35, 215]. This also tends to be true with haplochromine cichlids [216, 217]. Therefore, a laboratory experiment was set up with George Turner and his team in Bangor to test for reproductive isolation resulting from direct mate choice (Methods - section 6.8.4). Fifty Massoko females were given a choice from sixteen males representing the variety of male phenotypes. In parallel, I designed a SNP assay with 117 polymorphic sites representing 44 (HDRs) identified from the first 12 high-coverage Massoko samples sequenced in February 2013 (section 3.1). The regions are listed in Table 6.6.

Table 6.6 **Genotyped variants used for mate-choice trials and $\mathbf{F}_{ST}$ values observed in the reference sample of 18 benthic and 16 littoral males.** F_{ST} values are based only on the genotyped 18 Massoko benthic and 16 littoral males (the whole-genome sequenced individuals and other individuals used in the mate-choice trial are not included).

Variant coordinates	$\mathbf{F}_{ST}$	**Variant coordinates**	$\mathbf{F}_{ST}$	**Variant coordinates**	$\mathbf{F}_{ST}$
scaffold 1:4365113	0.466	scaffold 40:1588650	0.292	scaffold 88:1786645	0.71
scaffold 1:4365291	0.466	scaffold 40:1588728	0.292	scaffold 88:1786888	0.677
scaffold 6:7294300	0.405	scaffold 40:1621279	0.273	scaffold 88:1825810	0.71
scaffold 7:3305104	0.291	scaffold 40:1882810	0.206	scaffold 88:1825839	0.742
scaffold 7:3305216	0.291	scaffold 49:3025847	0.503	scaffold 88:1886989	0.665
scaffold 7:3313226	0.323	scaffold 55:2299953	0.025	scaffold 88:1923134	0.71
scaffold 7:3318131	0.262	scaffold 55:3423696	0.419	scaffold 88:1940222	0.581
scaffold 7:3318579	0.262	scaffold 55:3424572	0.456	scaffold 88:1940476	0.677
scaffold 12:3793589	0.416	scaffold 55:3435034	0.382	scaffold 88:1942123	0.581
scaffold 12:3793994	0.416	scaffold 55:3435507	0.382	scaffold 88:2222087	0.243
scaffold 14:3663857	0.424	scaffold 55:3451516	0.303	scaffold 88:2222122	0.243
scaffold 14:3669941	0.424	scaffold 55:3453112	0.303	scaffold 88:2429606	0.665
scaffold 14:4168852	0.232	scaffold 55:3480538	0.345	scaffold 88:2470678	0.345
scaffold 15:2959443	0.396	scaffold 55:3483480	0.378	scaffold 88:2473383	0.345
scaffold 15:2962256	0.135	scaffold 67:1282346	0.171	scaffold 88:2487048	0.135
scaffold 15:5455316	0.22	scaffold 67:1284312	0.171	scaffold 91:11230	0.101
scaffold 15:5458471	0.174	scaffold 67:1288981	0.219	scaffold 91:54791	0.159
scaffold 15:7238850	0.659	scaffold 82:2709101	0.076	scaffold 91:55547	0.159
scaffold 15:7251797	0.701	scaffold 82:2724117	0.076	scaffold 91:117365	0.092
scaffold 15:7252309	0.657	scaffold 82:2731927	0.098	scaffold 91:528794	0
scaffold 15:7254754	0.659	scaffold 87:112	0.159	scaffold 91:530091	0.008
scaffold 15:7269475	0.659	scaffold 87:5003	0.159	scaffold 97:188537	0.145
scaffold 15:7269758	0.701	scaffold 87:50289	0.098	scaffold 97:193146	0.118
scaffold 18:4359768	0.082	scaffold 87:51344	0.072	scaffold 97:193356	0.19
scaffold 18:4362575	0.038	scaffold 88:1185176	0.198	scaffold 97:2055581	0.295
scaffold 19:377749	0.049	scaffold 88:1198991	0.496	scaffold 126:32880	0.279
scaffold 26:1611369	0.038	scaffold 88:1199168	0.382	scaffold 126:38797	0.249
scaffold 29:3346390	0.011	scaffold 88:1213550	0.496	scaffold 126:1275145	0.264
scaffold 30:4055115	0.458	scaffold 88:1213711	0.462	scaffold 126:1284768	0.377
scaffold 30:6447512	0.194	scaffold 88:1312010	0.616	scaffold 146:1672851	0
scaffold 30:6448182	0.307	scaffold 88:1312055	0.616	scaffold 155:1053156	0.232
scaffold 30:6452026	0.281	scaffold 88:1441936	0.71	scaffold 217:517341	0.419
scaffold 31:426382	0	scaffold 88:1484291	0.71	scaffold 241:14395	0.03
scaffold 31:1867496	0.104	scaffold 88:1488398	0.71	scaffold 241:16987	0.162
scaffold 34:2235172	0.112	scaffold 88:1539583	0.452	scaffold 241:32682	0.165
scaffold 34:2250784	0.152	scaffold 88:1647616	0.387	scaffold 259:243418	0.355
scaffold 34:2802787	0.026	scaffold 88:1654621	0.387	scaffold 259:249905	0.387
scaffold 35:1693366	0.071	scaffold 88:1675293	0.387	scaffold 316:212810	0.456
scaffold 38:642672	0.388	scaffold 88:1781770	0.71	scaffold 316:213151	0.456

I genotyped a reference sample of 18 benthic and 16 littoral males, demonstrating that the SNP assay can reliably separate the ecomorphs along the first principal component (PC1) in PCA (Figure 6.18A, top). 78 out of the 117 sites replicated with high F_{ST} (≥ 0.2) in these additional reference males (Table 6.6). I then genotyped all females and males participating in the mate-choice experiments (Figure 6.18A, bottom) and calculated an average of the PC1 distances between each female and the males she mated with during the experiment, as assayed by microsatellite paternity analysis by Alexandra Tyers at University of Bangor. Richard Challis at University of Bangor then found that compared with expectation under random mating (Methods - section 6.8.4), females had a moderate, but significant ($P=4.3\times10^{-5}$, paired t-test), preference for mating with males genetically similar to themselves (i.e. close to them along PC1) (Figure 6.18B), demonstrating association between HDR variants and mate choice. Assortative mating by genotype was strong among females with positive (littoral) PC1 scores ($P=5.9\times10^{-9}$, paired t-test), while no assortative mating was detected among females with negative (benthic) PC1 scores (Figure 6.18B).

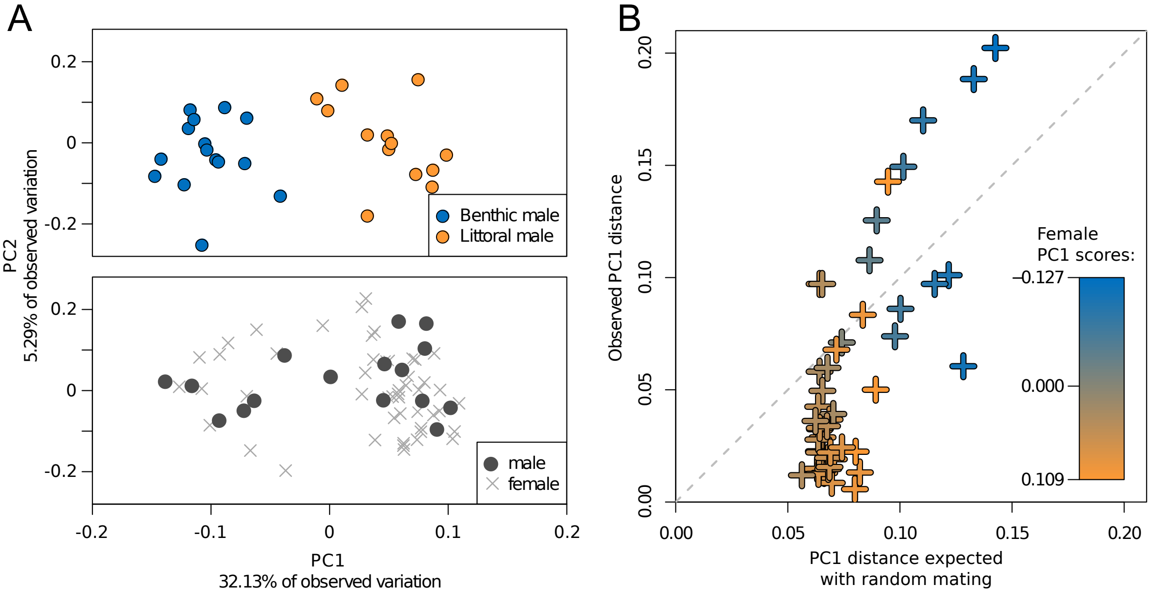


Fig. 6.18 **Mate-choice trials.** **(A)** PCA based on 117 genotyped SNPs. **Top:** The first axis of variation (PC1) in PCA reliably separates benthic and littoral males in a reference sample. **Bottom:** PC1 positions of females (N=50) and males (N=16) participating in mate-choice trials. **(B)** Results: Each point compares the average of absolute PC1 distances between a female and: males she mated with (observed PC1 distance) and all males she could have mated with (expected PC1 distance). Points are coloured according to the PC1 score of the female. Females below and to the right of the dashed diagonal line on average mate with males more like themselves in terms of PC1 score than would be true if they mated at random.

Stronger mating discrimination by ancestral populations compared to derived ones has been previously found in *Drosophila* and sticklebacks, possibly because low population density following a founder event favours less choosy individuals [218]. However, it is also possible that the benthic ecomorph only mates assortatively in the deep water environment; given that the experiments used wide-spectrum lighting characteristic of shallow water. Overall, the moderate assortative mating suggests a role for sexual selection in ecomorph divergence, but does not indicate that it is a primary force causing population-wide divergence.

6.6 Functions of adaptation

To explore the function of candidate adaptive genes, I performed Gene Ontology (GO) enrichment analysis [219]. Zebrafish (*Danio rerio*) has the most extensive functional gene annotation of any fish species. Therefore, I used the Broad Institute's assignment of orthologs between the *M. zebra* genome and zebrafish [93].

The numbers of genes available for GO analysis, genome-wide and in the enrichment gene-sets, are detailed in Table 6.7. Genome-wide, 13,230 (61.3%) of *M. zebra* genes have an assigned zebrafish ortholog, mapping to 11,810 unique zebrafish genes. Of these, ~7,000 (~33% of the total gene count) have useable GO annotation, the exact number depending on the GO category being assessed.

Table 6.7 **Numbers of genes available for Gene Ontology enrichment analysis** GO categories are denoted as follows: **`MF`** - molecular function; **`CC`** - cellular component; **`BP`** - biological process

Region	Total genes	Unique zebrafish orthologs	Orthologs with GO annotation		
			MF	CC	BP
Whole genome	21,567	11,810	7,086	6,441	6,961
Speciation islands ±50kb	215	146	73	72	75
All HDRs ±10kb	207	132	65	58	64
All HDRs ±50kb	398	288	135	123	133

GO enrichment analysis was performed on three enrichment gene-sets: a) genes in putative 'islands of speciation' ±50kb (enrichment terms in Table B.5); b) genes in all HDRs ±10kb (Table B.6); c) genes in all HDRs ±50kb (Table B.7). There is often an overlap between gene-sets annotated with different GO terms, in part because the terms are related to each other in a hierarchical structure [219]. Therefore, I used the Enrichment Map [220] app for Cytoscape (`http://www.cytoscape.org`) to organise all the significantly enriched terms into networks where terms are connected if they have a high overlap, i.e. if they share many genes. The resulting network, combining results of all three analyses, revealed clear clusters of enriched terms related to: a) morphogenesis (e.g. cartilage and pharyngeal system development, fin morphogenesis), consistent with morphological differentiation; b) sensory systems (e.g. photoreceptor cell differentiation), consistent with previous studies showing the role of cichlid vision in adaptation and speciation [89, 221]; and c) (steroid) hormone signalling (Figure 6.19).

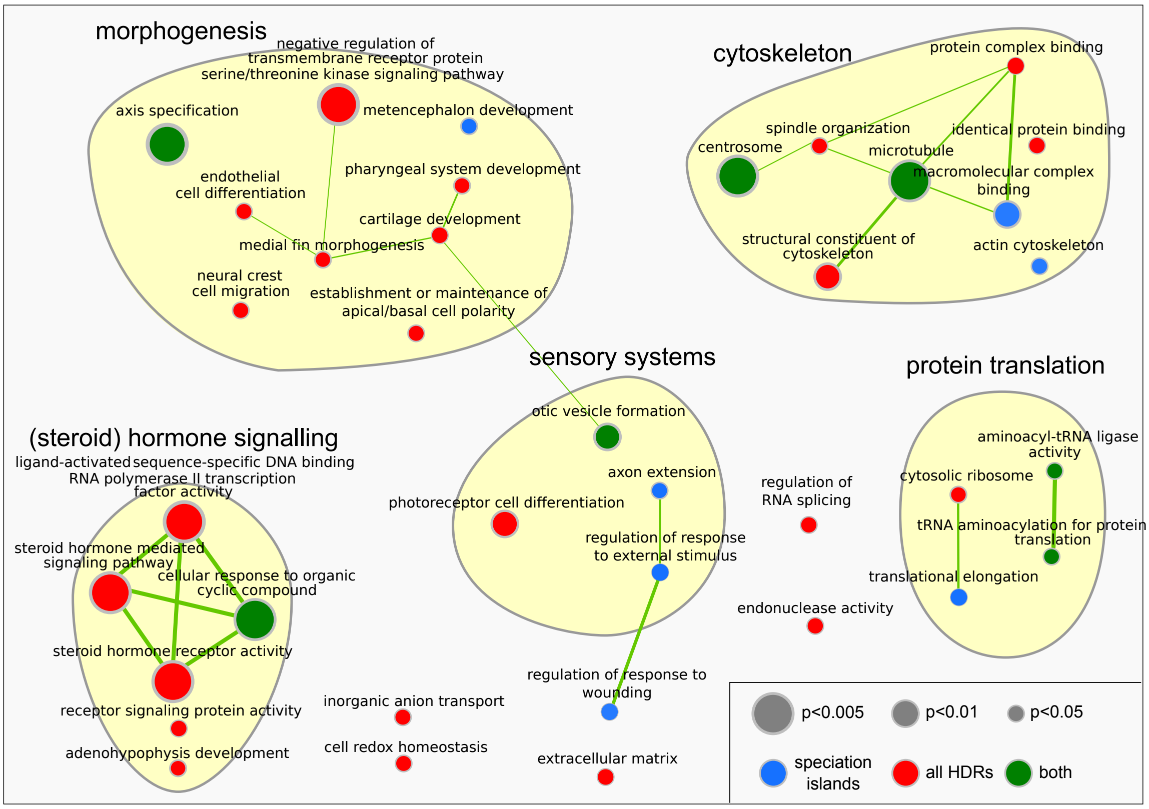
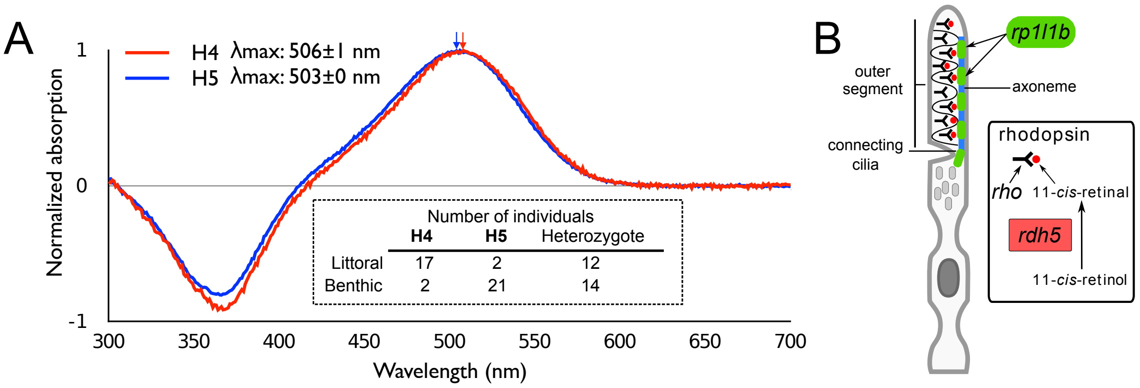

morphogenesis
negative regulation of transmembrane receptor protein serine/threonine kinase signaling pathway
axis specification
metencephalon development
endothelial cell differentiation
pharyngeal system development
cartilage development
medial fin morphogenesis
neural crest cell migration
establishment or maintenance of apical/basal cell polarity
cytoskeleton
protein complex binding
spindle organization
identical protein binding
centrosome
microtubule
macromolecular complex binding
structural constituent of cytoskeleton
actin cytoskeleton
sensory systems
otic vesicle formation
axon extension
photoreceptor cell differentiation
regulation of response to external stimulus
regulation of response to wounding
protein translation
aminoacyl-tRNA ligase activity
cytosolic ribosome
tRNA aminoacylation for protein translation
translational elongation
regulation of RNA splicing
endonuclease activity
(steroid) hormone signalling
ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity
steroid hormone mediated signaling pathway
cellular response to organic cyclic compound
steroid hormone receptor activity
receptor signaling protein activity
adenohypophysis development
inorganic anion transport
cell redox homeostasis
extracellular matrix
p<0.005
p<0.01
p<0.05
speciation islands
all HDRs
both

Fig. 6.19 **Enrichment Map for significantly enriched GO terms.** The level of overlap between GO enriched terms is indicated by the thickness of the edge between them. The size of the node indicates the best p-value for the term, and the colour of the node indicates the gene group for which the term was found significant (i.e. has P<0.05 in candidate ‘speciation islands’ ±50kb - blue; in all HDRs ±10kb or ±50kb - red; or in both groups - green). Broad functional groupings (morphogenesis, sensory systems...) were derived using automatic clustering using clusterMaker [222] and WordCloud [223], followed by manual editing.

The GO enrichment for terms related to sensory systems suggests a role of light environment heterogeneity in the divergence of the two forms. I examined in more detail the functions of candidate genes involved in photoreceptor function (Table 6.8), and two highly diverged alleles of the rhodopsin (*rho*) gene in Lake Massoko (alleles H4 and H5, separated by four amino acid changes; $F_{ST} = 0.39$; Figure B.4). Blue-shifted rhodopsin absorption spectra are known to play a role in deep-water adaptation [221].

Therefore, I initiated a collaboration with Yohey Terai (SOKENDAI, Japan) who expressed rhodopsins from H4 and H5 alleles, reconstructed them with 11-*cis*-retinal, and measured their absorption spectra (Methods - section 6.8.5). The results demonstrate that the H5 allele, associated with the deep-water benthic ecomorph, has a blue-shifted absorption spectrum (Figure 6.20A). The retina-specific retinol dehydrogenase *rdh5* (Table 6.8) produces 11-*cis*-retinal, the visual pigment binding partner of rhodopsin [224], and thus likely has a direct role in dark adaptation. Finally, a mouse ortholog of *rp1l1b* affects photosensitivity and morphogenesis of the outer segment (OS) of rod photoreceptor cells, locating to the axoneme of the OS and of the connecting cilia [225] (Figure 6.20B). Together, these results suggest divergent selection on *rho*, *rdh5*, and *rp1l1b* may facilitate the adaptation of scotopic (twilight) vision to the darker conditions experienced by the benthic ecomorph.



Number of individuals

	H4	H5	Heterozygote
Littoral	17	2	12
Benthic	2	21	14

Fig. 6.20 **Rhodopsin and rod cells. (A)** The absorption spectrum of the H5 allele of rho, more prevalent in the benthic ecomorph, is shifted towards blue wavelengths. **(B)** An illustration of the joint roles of *rho*, *rdh5*, and *rp1l1b* in photoreceptor rod cells. *rdh5* produces the chromophore 11-*cis*-retinal that binds *rho*, while *rp1l1b*, located at the axoneme of the outer segment and connecting cilia, also contributes to photosensitivity.

The role of the visual system in driving speciation along light gradients mediated by water depth has also been demonstrated in the *Pundamilia pundamilia* and *Pundamilia nyererei* cichlids of Lake Victoria [89]. In this system, the long-wavelength-sensitive opsin gene (LWS) variation is associated with water depth and male coloration, such that "more red-shifted alleles occur at greater depth" and "populations with predominantly red-shifted opsin alleles have predominantly red males". However, the LWS gene does not exhibit any protein-coding variation in Lake Massoko and it's location on scaffold 202 is not in proximity to Massoko HDRs.

Table 6.8 **Genes contributing to GO enriched terms in sensory perception**

sc	start	end	Gene symbol	Gene description	Entrez ID
15	5522881	5527067	rdh5	retinol dehydrogenase 5 (11-cis/9-cis)	556528
15	6719099	6756944	gnas	GNAS complex locus	557353
15	7474462	7493463	dctn2	dynactin 2 (p50)	394141
30	789904	799043	enpp4	ectonucleotide pyrophosphatase/phosphodiesterase 4	550586
30	853997	869417	rp1l1b	retinitis pigmentosa 1-like 1	101885561
57	14523	18913	mmp9	matrix metalloproteinase 9	406397
64	212132	217320	oprd1b	opioid receptor delta 1b	336529
66	676755	696647	bmper	BMP binding endothelial regulator	338246
84	2552689	2568793	chd	chordin	30161
99	277429	315316	cep290	centrosomal protein 290	560588
112	1985779	2057991	nsmfb	NMDA receptor synaptonuclear signaling and neuronal migration factor b	569891
164	58200	148583	plxnb1a	plexin b1a	561012

Several other HDR associated genes involved in vision have been studied in detail. The *dctn2* gene encodes the p50 subunit of the dynactin complex involved in microtubule-dependent intracellular transport. The strongest effects of *dctn2* morpholino knockdown in zebrafish were observed in observed in photoreceptor cells and in retinal radial glia. In addition, *dctn2* is involved in maintenance of the neuronal projection between the retina and the tectum of the midbrain, and in the survival of mechanosensory hair cells the in inner ear, and the lateral line [226]. A morpholino knockout of *cep290* revealed “defects in retinal, cerebellar, and otic cavity development” in zebrafish and immunogold labelling in mouse photoreceptor cells showed the greatest concentration in their connecting cilium, “supporting a possible ciliary role in the eye” [227].

Chordin and *bmper* both play key roles in many developmental processes by regulating bone morphogenesis proteins, including key roles in the formation of the otic vesicle (an embryonic structure that develops into the inner ear) [228, 229]. While the role of vision in general and opsin genes in particular has been studied extensively, these results suggest that fish divergence involving depth/light gradients may also involve hearing adaptations. The potential for hearing adaptation in fish divergence has so far been largely ignored, but it is likely that low light conditions at greater depths increase the importance of the sense of hearing, for example in predator/prey detection [230]. On the morphological side, hearing adaptations could be reflected for example in the size or shape of the otolith, or of the sensory hair cells [231].

The 14 genes driving GO enrichment in the morphogenesis functional grouping are listed in Table 6.9, and the 13 genes driving enrichment in steroid hormone signalling in Table 6.10. A detailed examination of the functions of these genes, possibly followed by further experiments, could shed light on the roles they may be playing in Massoko ecomorph divergence.

Table 6.9 **Genes contributing to GO enriched terms related to morphogenesis.**

sc	start	end	Gene symbol	Gene description	Entrez ID
15	7349839	7393790	gli1	GLI-Kruppel family member 1	352930
15	7474462	7493463	dctn2	dynactin 2 (p50)	394141
18	2784308	2814518	skib	nuclear oncoprotein skib	30113
18	3389782	3436376	sema3fa	semaphorin 3fa	544658
30	246439	259707	lypd6	LY6/PLAUR domain containing 6	447932
40	1538464	1598298	faf1	Fas associated factor 1	406243
57	1672259	1711225	rarga	retinoic acid receptor gamma a	30606
66	676755	696647	bmper	BMP binding endothelial regulator	338246
84	2552689	2568793	chd	chordin	30161
88	2438486	2477875	acvr2aa	activin A receptor type IIAa	553359
93	2203936	2236454	fbln1	fibulin 1	30240
93	2237233	2253337	fbln1	fibulin 1	30240
99	277429	315316	cep290	centrosomal protein 290	560588
114	1233887	1259975	fn1a	fibronectin 1a	100005469

Table 6.10 **Genes contributing to GO enriched terms in hormone signalling.**

sc	start	end	Gene symbol	Gene description	Entrez ID
0	14000583	14107901	rxraa	retinoid X receptor alpha a	555578
15	7349839	7393790	gli1	GLI-Kruppel family member 1	352930
15	7631396	7638449	tarbp2	TAR (HIV) RNA binding protein 2	336141
39	2242698	2298569	esr2b	estrogen receptor 2b	317733
40	1609941	1613581	dmrta2	doublesex and mab-3 related transcription factor like family A2	474350
57	1658903	1665415	nr1d4a	nuclear receptor subfamily 1 group D member 4a	563150
57	1672259	1711225	rarga	retinoic acid receptor gamma a	30606
64	77011	78338	nr0b2a	nuclear receptor subfamily 0 group B member 2a	403010
74	637448	647727	vtg3	vitellogenin 3 phosvitinless	30518
88	778189	875341	ahr2	aryl hydrocarbon receptor 2	30517
88	2438486	2477875	acvr2aa	activin A receptor type IIAa	553359
108	2974	18121	zgc:171775	zgc:171775	562552
186	675492	721013	rgs12b	regulator of G-protein signaling 12b	378970

6.7 Comparisons to other systems

Overall, the results presented in this chapter suggest a pair of incipient species undergoing divergence with gene flow within the crater lake Massoko. Their overall level of divergence (F_{ST} = 0.038) is low compared with background F_{ST} observed in other recent studies of speciation with gene flow in *Anopheles* mosquitoes (S and M form; F_{ST} = 0.21) [66], *Ficedula* flycatchers (F_{ST} = 0.36) [70], and *Heliconius* butterflies (F_{ST} = 0.18) [69], highlighting that we are looking at an early stage of divergence. The MSMC analysis suggests that median effective divergence occurred within the last 500-1,000 years (~200-350 generations), following separation of lake fish from the Mbaka river population around 10,000 years ago (Fig. S4). However, divergence may have started considerably earlier than these times, masked by subsequent gene flow.

Among populations at similar levels of divergence to Lake Massoko ecomorphs are *Timema* stick insects (F_{ST} = 0.015 for adjacent and F_{ST} = 0.03 for geographically isolated population pairs), where thousands of regions of moderately elevated divergence were found all across the genome [232], and German carrion and Swedish hooded crows (F_{ST} = 0.017), that have strongly diverged with fixed differences, but at fewer than five loci [77]. In Massoko, we observe an intermediate pattern between these two extremes, with a few dozen moderately elevated islands, clustering within the genome indicating close linkage, and no fixed differences. A genome-wide pattern with multiple loci of moderate divergence suggests a genomic architecture similar to the ecological divergence of a sympatric threespine stickleback pair in Paxton Lake, Canada [233], and the sympatric divergence of dune-specialist sunflowers, Helianthus [234].

6.8 Detailed methods

6.8.1 Field sampling and eco-morphological analysis

Field sampling for genetic, morphological and stable isotope samples

Astatotilapia samples from Lake Massoko were collected on 17th July 2011, and from 19th to 25th November 2011 by Martin Genner, George Turner and their teams. Fish were collected using fixed gill nets and SCUBA. On being brought to the surface, fish were given an overdose of anaesthetic (MS-222). From each fish we collected a genetic sample (fin clip) that was stored in ethanol, and cut a fillet of the flank for stable isotope analyses that was sun-dried and stored with desiccant. Samples of potential food sources were also collected and dried, including epilithic algae, sponges and bivalves. Whole fish were preserved in formalin (~4%). Genetic samples (fin clips) of outgroup *Astatotilapia calliptera* were collected opportunistically between 2009 and 2014.

Field sampling for assessment of ecomorph frequency with depth

Astatotilapia samples from Lake Massoko were collected from 28th July to 7th August 2014, 10th to 15th December 2014, and 5th - 24th August 2015. Fish were collected using fixed gill nets, angling and SCUBA. I was part of the field team during the first two trips, in charge of the gill net sampling. Depth was assessed using a plumbline, surface depth meter, and dive gauges. Fish were photographed on collection. All adult males >65mm Standard Length were assigned to an ecomorph on the basis of the colour of body and fins, and gross morphology. Depth was recorded as bottom depth, which means that fish caught in the water column may sometimes be included. This, along with drift of passive fishing gears may perhaps have led to an over-representation of shallow water fish in deeper water records

6.8.2 RAD-seq data processing and analysis

Note: The work on RAD-seq data described in this section was done by Richard Challis, University of Bangor. The methods are included here (in this thesis) for completeness.

DNA extraction and sequencing

DNA was extracted from ethanol-preserved fin tissue from 56 wild caught fish using a standard CTAB-Chloroform extraction method including an RNAase treatment step. This was sent to Floragenex (`http://www.floragenex.com/`) for library preparation using the Sbf1 enzyme and sequencing on an Illumina HiSeq2000 platform, providing 100bp single end reads. The samples were sequenced in two rounds. In the first round sequencing was 28 samples per lane, but 43 individuals obtained less than 1M reads each. In a second round 41 of these 43 individuals were reprepped, and sequenced at 20 and 21 samples per lane. Raw data have been deposited at the NCBI Sequence Read Archive under BioProject PRJNA286304 (Accessions SAMN03768857 to SAMN03768912).

Variant calling and filtering

Samples with fewer than 300000 reads (approximately 20X coverage per tag) were removed. Raw reads for the remaining 42 samples were demultiplexed and adaptor trimmed leaving 89 base reads for use in reference guided RAD tag analysis. Reads were aligned to the Mbaka River *Astatotilapia calliptera* consensus sequence (see section 6.8.3) using `bwa-mem v0.7.12` [132]. An average of 96.2% (±0.3%) of reads mapped to the reference and these mapped reads were filtered to remove reads with terminal alignments and reads that were not uniquely mappable leaving an average of 90.3% (±1.1%) of the original reads in the filtered read set. SNPs were called using the stacks [235] `ref_map.pl` pipeline with a minimum stack depth (`-m`) of 5. The full dataset was filtered to remove SNPs that had been called in less than 75% of samples and the resulting matrix contained 7,906 SNPs and was 82.3% complete.

Phylogenetic trees and constraint tests

Phylogenetic model testing using `ModelGenerator v0.85` [236] supported the use of the GTR + Γ model of sequence evolution with an estimated transition/transversion ratio of 2.65. A maximum likelihood (ML) phylogeny was produced using `RAxML v8.0.22` [198] using the GTRGAMMA model. Support for the ML tree topology was inferred using 100 rapid bootstrap samples [199]. The phylogeny was rooted on *A. tweddlei* and has been deposited in treebase (accession: TB2:S18241).

6.8.3 Whole genome data processing and analysis

Gene annotation

For analyses concerning gene content and function, I used the `V1` gene annotations generated at the Broad Institute as a part of the cichlid genome project [93]. The includes assignment of orthologs between cichlid and medaka, tetraodon, stickleback, and zebrafish genomes.

Whole genome phylogenetic tree

Consensus genome sequences were generated using the `bcftools v1.2 consensus` tool. For each sample, I selected the sequence of one haplotype (as assigned by `beagle` haplotype phasing - see section 3.2.4) by using the `--haplotype=1` option in `bcftools`. All scaffolds except the mtDNA sequence (scaffolds 747, 2036) were concatenated into a single sequence and phylogenetic trees then inferred using `RAxML v7.7.8` [198] under the GTRGAMMA model (General Time Reversible model of nucleotide substitution with the Γ model of rate heterogeneity). The maximum likelihood tree was obtained as the best out of five alternative runs on distinct starting maximum parsimony trees (using the `-N 5` option). Sixty six bootstrap replicates were obtained using `RAxML's` rapid bootstrapping algorithm [199]. It was my intention to run more bootstrap replicates, enough to satisfy `RAxML -N autoFC` frequency-based bootstrap stopping criterion, but this has proven computationally infeasible on a dataset of this size (obtaining the 66 replicates required ~7,647 hours of CPU time). Still 66 replicates provide a reasonable indication of bootstrap support for the maximum likelihood tree. Bipartition bootstrap support was drawn on the maximum likelihood tree using `RAxML -f b` option.

Principal Component Analysis

SNP variants (no indels) with minor allele frequency >= 0.05 were selected using `vcftools v0.1.12b` options `--remove-indels --maf 0.05` and exported in `PLINK` format [194]. The variants were LD-pruned to obtain a set of variants in approximate linkage equilibrium (unlinked sites) using the `--indep-pairwise 50 5 0.2` option in `PLINK v1.0.7`. Principal Component Analysis on the resulting set of variants was performed using the `smartpca` program from the `eigensoft v5.0.1` software package [195] with default parameters.

Linkage Disequilibrium

First, I obtained a random subsample of approximately 10% of all SNP variants (no indels) from the full joint Massoko, Itamba and *Astatotilapia calliptera* variant set. Then, linkage disequilibrium (LD) was calculated for each variant within 1Mbp window using `vcftools v0.1.12b` options `--hap-r2 --ld-window 1000000`. The plots use the R^2 measure of LD, and show averages within 1000bp windows calculated in the `R` software environment for statistical computing [175].

ADMIXTURE ancestry estimation

All SNP variants were exported in PLINK format [194] and LD-pruned to obtain a set of variants in approximate linkage equilibrium (unlinked sites) using the `--indep-pairwise 50 5 0.2` option in `PLINK v1.0.7`. The `ADMIXTURE v1.23` program was then run with default parameters. The postulated number of ancestral populations `K` was set to 1, 2, 3, 4, 5, and 6. From a statistical standpoint, the authors of the software suggest choosing the value of `K` with the lowest cross-validation error. We performed 10-fold cross-validation (`--cv=10`) and found the lowest cross-validation error is with `K=1` (Figure 6.4). ADMIXTURE cross-validation implying `K=1` is a common phenomenon when population differentiation is subtle, but meaningful results can still be obtained with higher values of `K` - see for example the application of ADMIXTURE to HGDP human European data in [193] and Figure S12 therein.

Accessible genome

To obtain an estimate of the length of the genome accessible for accurate variant calling using Illumina short reads I used Heng Li's `SNPable` tool (available from `http://lh3lh3.users.sourceforge.net/snpable.shtml`). The SNPable tool divides the reference genome into overlapping *k*-mers (sequences of length *k* - I used *k*=50), and then the extracted *k*-mers are aligned back to the genome (I used `bwa aln -R 1000000 -O 3 -E 3`). Then I only kept regions where the majority of overlapping 50-mers were mapped back uniquely and without 1-difference. Excluding gaps (runs of `N` in the reference), this SNPable 'mask' excludes approximately 7.4% of the reference, resulting in an 'accessible genome' of 660,796,086bp.

Heterozygosity

The number of heterozygous sites per individual was calculated using the custom C++ program `evo` (available from `https://github.com/millanek/evo`), with the `stats`

`--hets-per-individual` option, and then divided by the length of the 'accessible genome' (see above) to obtain the frequency of heterozygous sites.

MSMC cross-coalescence analysis

Because results of MSMC rely, in part, on detecting the density of heterozygous sites, we restricted this analysis to high coverage (~15X) samples. Genomic regions on which short reads cannot be uniquely mapped were masked out by a) excluding genomic regions where mapped depth was higher than 35X (more than twice the average genome coverage); b) using Heng Li's SNPable tool (see 'accessible genome' above).

Running MSMC without the `-fixedRecombination` parameter for 100 iterations indicated that the rho/mu parameter is approximately 2. This value was used for all following MSMC runs (i.e rho/mu parameter set to 2: `msmc --rhoOverMu=2 --fixedRecombination -P 0,0,1,1`). Each run of the cross-coalescence analysis used four haplotypes, two from each ecomorph for the benthic-littoral split, and two from Mbaka and two from Massoko for the Massoko-Mbaka split.

Since MSMC relies on long-range haplotype phasing, we re-phased the data using the `shapeit v2.r790` haplotype phasing method [237] including the use of phase-informative reads [238]. Because of the need for long-range phase information, we restricted the analysis to 50 largest genomic scaffolds, comprising ~390Mb of sequence.

MSMC N_e history estimation

The analysis was performed as the above cross-coalescence runs but without the `-P` parameter (i.e. `msmc --rhoOverMu=2 --fixedRecombination`), except that low coverage (~6X) samples had to be used for Lake Itamba. For these samples, I excluded (masked out) genomic regions where mapped depth was higher than 20X.

Coalescent simulations

I used the coalescent simulator `ms` [211] to simulate the divergence of two subpopulations, sampling 74 chromosomes from the first population corresponding to 37 *Massoko benthic* samples and 64 chromosomes from the second population corresponding to 32 *Massoko littoral* samples (`-I 2 74 64` parameter). The simulations were performed under a range of models and demographic scenarios, as described in the main text. Migration rate for the Isolation with migration (IWM) model was included directly in the `-I` parameter (e.g. `-I 2 74 64 5` and for the Isolation after migration migration model was adjusted using the

For each model/scenario: a) the between-population split time (`-ej` parameter) was adjusted to match the overall observed benthic littoral F_{ST} of 3.89%; b) I simulated 500,000 independent samples, each sample with one segregating site (effectively simulating 500 thousand unlinked loci). Therefore, the basic command line for the IWM model looked as follows:

```
ms 138 500000 -s 1 -I 2 74 64 M -ej splitT 1 2
```

where `M` is the migration parameter (see Table B.3), and `splitT` stands for the split time.

Calculating F_{ST} and defining HDRs

F_{ST} was calculated both for simulations and for the cichlid data using my own code implemented in the C++ program `evo` (available from `https://github.com/millanek/evo`), with the `fst --ms` option for simulations and `fst --vcf` option for cichlid data. The F_{ST} calculation implements the Hudson estimator, as defined by Bhatia, Patterson *et al.* [197, equation 10], using 'ratio of averages' to combine estimates of F_{ST} across multiple variants, as recommended in their manuscript.

For defining HDRs, I used windows of 15 variants each, which I found to provide good balance between fine genomic resolution and reducing stochastic variation by averaging over variants. Nevertheless, I found some cases where F_{ST} between neighbouring regions dipped briefly below the threshold, which I believe to be in most cases due to remaining stochastic variation. The length (the extent) of HDRs was defined by merging windows with $F_{ST} >= 0.25$ that were next to each other or within 10,000bp of one another using `bedtools v2.16.2` [239]: `mergeBed -d 10000 -i windows_fst_above0.2.bed`. F_{ST} was also calculated in 10kb windows and each HDR must contain at least one window with $F_{ST} >= 0.3$, as described in the main text.

Characterisation of HDRs in terms of d_{XY}, and π_{diff}

Both d_{XY} and nucleotide diversity (π) were calculated for 10kb windows. The d_{XY} statistic was calculated as defined by Wakeley [240, equation 3]. Both calculations are implemented in my C++ program `evo` (available from `https://github.com/millanek/evo`), and were obtained by using the `fst --vcf` option.

Average nucleotide diversity in each window was calculated separately for the benthic (π_B) and littoral (π_L) ecomorphs and π_{diff} was then calculated as the absolute value of the difference between π_B and π_L; i.e. $\pi_{diff} = |\pi_B - \pi_L|$. The 'direction' of the 'sweep' is in the morph with lower π; i.e. if $\pi_B < \pi_L$ then the potential 'sweep' was inferred to be in the benthic morph.

Gene Ontology enrichment analysis

Gene Ontology (GO) enrichment for genes found within HDRs was calculated in `R` [175] using the `topGO` package [241] from the Bioconductor project [242]. The GO hierarchical structure was obtained from the `GO.db` annotation [243] and linking zebrafish gene identifiers to GO terms was accomplished using the `org.Dr.eg.db` annotation package [244].

Chromopainter and fineSTRUCTURE

Singleton SNPs were excluded using `bcftools-1.1 -c 2:minor` option, before exporting the remaining variants in PLINK format [194]. The `chromopainter v0.0.4` software [193] was then run for 150 largest genomic scaffolds. Briefly, I created a uniform recombination map using the `makeuniformrecfile.pl` script, then estimated the effective population size (N_e) for a subsample of 20 individuals using the chromopainter inbuilt expectation-maximization procedure [193], averaged over the 20 N_e values using the provided `neaverage.pl` script. Estimated N_e values ranged from 1,046 to 6,015 (mean 3914, sd. 990). The chromopainter program was then run for each scaffold independently, with the `-a 0 0` option to run all individuals against all others. Results for individual scaffolds were combined using the `chromocombine` tool before running `fineSTRUCTURE v0.0.5` with 1,000,000 burn in iterations, and 200,000 sample iterations, recording a sample every 1,000 iterations (options `-x 1000000 -y 200000 -z 1000`). Finally, the sample relationship tree was built with fineSTRUCTURE using the `-m T` option and 20,000 iterations.

Patterson's D (ABBA-BABA) and related statistics

To test for possible gene-flow between surrounding rivers and Massoko, I calculated the ABBA-BABA statistic (16, 17). The ABBA-BABA test (also known as *D statistic* or *Patterson's D*) tests for introgression an excess of shared derived alleles between one of two populations and an outgroup. Formally, I calculated D(benthic, littoral, Mbaka river, *P. nyererei*) using equation S15.2 of Green et al. [190], allowing me to use allele-frequency information from all benthic and littoral individuals. I also estimated *f*, the admixture fraction following Green *et al.* equation S18.5, and calculated the standard error for both estimates by a weighted block jackknife, using blocks of 5,000 informative variants (i.e. variants with ABBA or BABA patterns).

The D and f statistics were calculated genome-wide and D and f_{dM} also in non-overlapping windows of 50 informative variants each. To add ancestral allele information (i.e. the outgroup variants) to the crater lakes VCF file, I used whole genome alignment between *M. zebra* and *P. nyererei*, as described in section 3.4. The f_{dM} statistic has been calculated as defined previously in section 5.5.

6.8.4 Mate choice experiments

Note: The mate choice trials were done by Alexandra Tyers, University of Bangor. The methods are included here (in this thesis) for completeness. As also noted in the main text, I designed the SNP assay that formed the basis for similarity scoring during the analysis of these trials, and contributed to the analysis.

Experimental setup, and aquarium work

A single 4m long tank with a gravel/sand substrate was divided into eight sections by ‘partial partition’ grids [245]. Each section contained a terracotta plant pot (to function as territorial focal points for the males) and a selection of plastic plants. Water was filtered and heated (to ~26 C) externally and the tank lit from above by white and UV enhanced fluorescent tube lamps. Fish were fed daily with algae flake and 2-3 times weekly with frozen bloodworm.

Two female mate choice trials were carried out using two different sets of eight males and a total of 50 females. All fish were wild caught and shipped to the UK in December 2011. Trial 1 ran from the beginning of November 2012 to the end of January 2013 (3 months) and trial 2 started at the beginning of February 2013, ending in June 2013 (4.5 months). Each set/trial comprised 3-4 large littoral, 1-2 large benthic and 3-4 ‘small’ males. Forty-five of the 50 females produced broods in both trials. All of the larger littoral and benthic males within each set were of a comparable size and unable to fit through the partial partition grids. The large males were placed in every-other section, leaving the territories in-between available to the small males which, being of a similar size to some of the larger females, were also able to move freely between sections. Before introduction of the females to the experimental tank, males were left until the smaller ones had settled into the ‘empty’ territories between the bigger males.

As with other haplochromine cichlid fish, *Astatotilapia* are maternal mouth-brooders, egg are picked up by the female during spawning and protected in the buccal cavity during development before release as free-swimming young approximately three weeks later. Females were removed from the experimental tank after spawning and isolated in small tanks on a recirculating system during the brooding phase. After the first trial, offspring were gently removed from the females mouths after 10 days and euthanised by anaesthetic (clove oil) overdose. Females were kept in their individual tanks to allow for rest and recovery before the second trial. Once all females had spawned in the first trial, the males were changed. Allele diversity at the chosen microsatellite loci (Ppun5, 7 & 21) [246] was sufficiently high to allow for the identification of all individual females by their microsatellite profile, it was therefore possible to return all females to the experimental tank at the same time for the second trial and re-identify individuals later during the second round of paternity testing. After spawning in the second trial, females were again isolated, but left to brood to term. Five offspring from each brood were euthanised for paternity testing.

Paternity testing

475 offspring from 95 broods (five per brood/trial), produced over the two replicates were genotyped for paternity analysis (250 from 50 broods in trial 1; 225 from 45 broods in trial 2). Tissue was taken from ethanol preserved fry samples and DNA obtained by salt extraction. DNA samples from offspring, mothers, and all potential fathers were used for assigning paternity by allele sizing after PCR multiplex (Qiagen multiplex kit) of three microsatellite markers (Ppun5, 7 & 21) [246]. Genotyping of the amplified samples was carried out on an Applied Biosystems (ABI) 3130xl genetic analyzer using LIZ 500(-250) (ABI) size standard. The genotype of each individual (males, females, offspring) were determined by manual scoring of alleles in `Peak Scanner v2`. 447 (94%) of the genotyped offspring were successfully assigned to an individual male. Due to allele sharing among males used in trial 2, 23 offspring could not be assigned unambiguously. Seven offspring could not be assigned due to problems with amplification or disagreement between microsatellite loci (possible cross-contamination). Overall, 8-10 offspring from each female that produced more than one brood, were unambiguously assigned to father.

Forty-six of the 50 females spawned with more than one male during the course of the experiment and some females were found to have spawned with up to four males in total (44% of individual broods were sired by more than one male).

Data analysis

I designed a Sequenom MassARRAY SNP genotyping assay [247] for 117 SNPs, over four Sequenom plates. The assay was performed by the Wellcome Trust Sanger Institute core genotyping team. Analysis of the SNP data was performed in `R` using the `Bioconductor` [248] package `SNPRelate` [249] to account for linkage disequilibrium (LD) between the SNPs. SNPs were filtered using a recursive sliding window approach (`snpgdsLDpruning`) with an LD threshold of 0.2. Principal components analysis of the filtered dataset was used to obtain a score for each of the individuals used in the mate choice experiments.

The expected distance between female and male PC1 scores (under the null hypothesis of no assortative mating) was calculated as follows:
Expected value = mean absolute distance between female PC1 score and PC1 scores all the possible combinations of males she might have mated with.
The **observed value** is the mean absolute distance between female PC1 score and PC1 score of all males mated with.

The above calculations are based on the total number of males a female actually mated with and the number of trials she took part in. The values are therefore different for each female because some did not take part in both trials (or did not spawn in both trials), and there was variation in the total number of males mated with over the course of the experiment (between 2-4).

For each female, the number of potential mates is a product of the possible combinations in trial 1 (T1) and trial 2 (T2). The number of combinations (choosing r males out of n males) in each trial are n!/r!(n-r)!. For example, the number of combinations is 7 for a female that mated with a single male in T1 and 588 for a female that mated with 2 males in T1 and 2 males in T2.

6.8.5 Measuring rhodopsin absorption spectra

Note: The work described here was done by Yohey Terai, SOKENDAI, Japan, starting with DNA samples sent by myself. The methods are included here (in this thesis) for completeness.

Reconstruction and measurement of absorption spectra of visual pigments

Production, reconstruction, purification, and measurement of the visual pigments were performed as described Ueyama *et al.* [250] with minor modifications. Briefly, the sequences of *rho* (also known as RH1) H4 and H5 alleles were amplified by PCR using genomic DNA of Lake Massoko cichlids as a template with a pair of specific PCR primers [221] designed to produce a fusion protein with a FLAG-tag (Sigma-Aldrich) at its C terminus. The amplified DNA fragments were digested with restriction enzymes and cloned into the expression vector pFLAG-CMV-5a (Sigma-Aldrich). The visual pigments were reconstituted with A1-derived retinal. Absorption spectra of the pigment solutions in the presence of hydroxyl-amine (<100mM) before and after photobleaching were recorded using a spectrophotometer (UV-2400, Shimadzu, Japan). The measurements were taken 5 times before and after photobleaching. We determined the mean peak spectral values (maximum absorption spectra: λmax) and standard errors from multiple preparations and measurements for each pigment. All procedures after reconstitution of the pigments were performed under dim red light (>680 nm) conditions.

Chapter 7

Conclusions

7.1 Conclusions

In this thesis, I describe the some of the first steps in the application of whole genome sequence data to study the exceptional diversity of East African cichlids. I generated an annotation of microRNA loci for five reference genomes, predicted which protein coding genes may be regulated by them, and explored sequence evolution both in the genes themselves and their targets. Then I took advantage of one of the reference genomes (the *M. zebra* genome for Lake Malawi) and aligned whole genome sequences from 239 individuals to it to obtain two detailed catalogues of variation: over 20 million genetic variants for Lake Malawi and almost 5 million variants to study incipient speciation of *Astatiotilapia* cichlids in the isolated crater lake Massoko. Another 29 Lake Malawi and 4 crater lake region *Astatotilapia* individuals have been sequenced but their genomes have not yet been analysed. To further facilitate research on this fascinating system, I constructed whole genome alignments between the reference genome assemblies, assigned ancestral alleles to genetic variants in Lake Malawi, and built a genome browser to visualise genome wide datasets for East African cichlids.

During an initial analysis of the Lake Malawi dataset I found that heterozygosity based N_e estimates range from ~7,200 to ~24,300, and that F_{ST} between Lake Malawi species varies between 0.04 and 0.66. A phylogenetic approach to the study of species relationships provides a strong signal when averaging across all genomic loci, but there is a lot of discordance between local phylogenies, and especially between phylogenies built using nuclear and mitochondrial DNA data. The discordance is due to high prevalence of incomplete lineage sorting, but interspecific introgression may also play a role. In three cases of possible introgression I formally tested the hypothesis using the ABBA-BABA statistic. In the first case, where introgression was suggested by